SHARPNESS AND CONDITIONING OF NONSMOOTH CONVEX FORMULATIONS IN STATISTICAL SIGNAL RECOVERY*

LIJUN DING† AND ALEX L. WANG‡

Abstract. We study a sample complexity vs. conditioning tradeoff in modern signal recovery problems (including sparse recovery, low-rank matrix sensing, covariance estimation, and abstract phase retrieval), where convex optimization problems are built from sampled observations. We begin by introducing a set of condition numbers related to sharpness in the ℓ_1 or Schatten-1 norm of nonsmooth formulations for these problems. Then, we show that these condition numbers become dimension *independent* constants in each of the example signal recovery problems once the sample size exceeds some constant multiple of the recovery threshold. Structurally, this result ensures that the inaccuracy in the recovered signal due to both observation noise and optimization error is controlled. Algorithmically, such a result ensures that a new restarted mirror descent method achieves nearly-dimension-independent linear convergence to the signal in terms of iterations. This new first-order method is general and applies to any sharp convex function in an ℓ_p or Schatten-p norm for $p \in [1, 2]$.

Key words. Sharp formulation, signal recovery, restarted mirror descent, sparse recovery, matrix sensing, phase retrieval

MSC codes. 90C30, 90C06, 90C22

1. Introduction.

This paper studies a *sample complexity vs. convex optimization conditioning* tradeoff in modern signal recovery problems such as sparse recovery [15], low-rank matrix sensing [9, 44], phase retrieval [14], and covariance estimation [20]. To illustrate the setup, the challenges, and our results concretely, we will only consider the abstract phase retrieval problem (henceforth phase retrieval) in the introduction. The general setup and results follow a similar pattern and are deferred to later sections.

Phase retrieval asks us to recover an unknown rank-one positive semidefinite (PSD) matrix $X^\natural \in \mathbb{S}^n$ given a vector of m measurements $b = \mathcal{A}(X^\natural)$ where the linear map $\mathcal{A} : \mathbb{S}^n \to \mathbb{R}^m$ is defined by $\mathcal{A}(X)_i = g_i^\intercal X g_i$ for i.i.d. $g_i \sim N(0, I_n/m)$. Here and throughout, $\mathbb{S}^n$ and $\mathbb{S}^n_+$ denote the vector space of symmetric matrices and the cone of PSD matrices respectively. Concretely, we are asked to

$$\text{recover } X^\natural \in \mathbb{S}^n \text{ from } (b, \mathcal{A}) \text{ where } b = \mathcal{A}(X^\natural) \in \mathbb{R}^m.$$

In [10], Candès and Li established an *exact recovery* result: If $m = \Theta(n)$, then, with high probability (w.h.p.)[1], $X^\natural$ is the unique optimizer for the convex optimization problem

$$\min_{X \in \mathbb{S}^n} \left\{ \operatorname{tr}(X) : \begin{array}{l} \mathcal{A}(X) = b \\ X \in \mathbb{S}^n_+ \end{array} \right\}. \tag{1.1}$$

Here, the choice of the objective function $\operatorname{tr}(X)$ and the constraint $X \in \mathbb{S}^n_+$ reflect prior knowledge that the true signal is low rank and PSD respectively.

*Submitted to the editors January 25, 2026.

Funding: This work was completed in part while A. L. Wang was supported by the Dutch Scientific Council (NWO) grant OCENW.GROOT.2019.015 (OPTIMAL). L. Ding was supported by an NSF TRIPODS grant to the Institute for Foundations of Data Science (NSF DMS-2023239).

†Department of Mathematics, University of California San Diego, La Jolla, CA (l2ding@ucsd.edu, https://www.lijunding.net/).

‡Daniels School of Business, Purdue University, West Lafayette, IN (wang5984@purdue.edu, https://web.ics.purdue.edu/~wang5984/)

[1] An event happens w.h.p. if the probability is larger than $1 - c_1 \exp(-c_2 m)$ for some numerical constants c_1, c_2.

Issues and challenges. Unfortunately, convexity and uniqueness of solution are not the end of the story as (1.1) may not be well-conditioned. In such a situation, the recovery process may suffer from:

- *measurement error*: if the observation b is corrupted by noise, the optimal solution of (1.1) may deviate greatly from $X^\natural$ or may even cease to exist;
- *optimization error*: the problem (1.1) may admit near-optimal and near-feasible solutions in terms of function value and constraint violation that are far from the true signal $X^\natural$;
- *slow algorithm convergence*: the iteration complexity of existing first-order algorithms for solving (1.1) may depend polynomially on the dimension due to polynomially poor conditioning of (1.1).[2] Such dependence weakens the applicability of existing first-order methods on large-scale instances of (1.1).

Our contribution. In this paper, we study a notion of conditioning related to *sharpness* for (1.1) based on an *unconstrained nonsmooth reformulation* of (1.1). This notion of conditioning is measured in the Schatten-1 norm and allows us to quantitatively control each of the above issues. To measure the conditioning of (1.1), we introduce the following penalized nonsmooth version of (1.1):

$$\min_{X\in\mathbb{S}^n} F_{r,\ell}(X) \coloneqq \operatorname{tr}(X) + r\left\|\mathcal{A}(X)-b\right\|_1 + \ell \operatorname{dist}_1(X,\mathbb{S}^n_+). \tag{1.2}$$

Here, the distance $\operatorname{dist}_1(X,\mathbb{S}^n_+) \coloneqq \inf_{Y\in\mathbb{S}^n_+}\|Y-X\|_1$ is measured in the Schatten-1 norm, i.e., the sum of the singular values, and $r,\ell\geq 0$ are penalty parameters on the violations of $\mathcal{A}(X)=b$ and $X\in\mathbb{S}^n_+$ respectively. To clean up the exposition, we will understand "ℓ_p norm" of a matrix to mean its Schatten-p norm and "ℓ_p norm" of a vector to mean its usual ℓ_p norm. We will also abuse terminology and refer to the rank of a matrix or the size of the support of a vector as its "ℓ_0 norm". We emphasize that we penalize the violation of $\mathcal{A}(X)=b$ by $\|\mathcal{A}(X)-b\|_1$ and not the more common least-squares error $\|\mathcal{A}(X)-b\|_2^2$ and we use the ℓ_1 norm in the distance function. The term $\|\mathcal{A}(X)-b\|_1$ is sometimes referred to as the least absolute deviation error term [43] and is used as a robust version of the least-squares error. Although this choice leads to a nonsmooth problem, it also opens the possibility for the function (1.2) to be μ-sharp around $X^\natural$. Specifically, we will show that $F_{r,\ell}$ is μ-sharp around $X^\natural$ w.r.t. the ℓ_1 norm, that is, for all $X\in\mathbb{S}^n$

$$\underbrace{\operatorname{tr}(X) + r\left\|\mathcal{A}(X)-b\right\|_1 + \ell \operatorname{dist}_1(X,\mathbb{S}^n_+)}_{F_{r,\ell}(X)} - \underbrace{\operatorname{tr}(X^\natural)}_{F_{r,\ell}(X^\natural)} \geq \mu\left\|X-X^\natural\right\|_1. \tag{1.3}$$

When such a bound holds, we will think of the parameters (μ,r,ℓ) as partially describing the conditioning of (1.1). Finally, in order to make our notion of conditioning "aware" of scaling, we additionally track the Lipschitz constant, L, of (1.2) with respect to the ℓ_1 norm:

$$|F_{r,\ell}(X)-F_{r,\ell}(Y)| \leq L\left\|X-Y\right\|_1, \quad \text{for all } X,Y\in\mathbb{S}^n. \tag{1.4}$$

We use the set of numbers (μ,r,ℓ,L) to measure the conditioning of (1.1). The sharpness parameter μ in some sense will control absolute notions of error, while the

[2]These polynomial factors may be even worse when measuring error in terms of distance to the true solution as opposed to function value.

condition number $\kappa := L/\mu$ will control relative notions of error and convergence rates of first-order methods.[3]

We now describe our main results in the setting of phase retrieval:

- *Conditioning*: It is well-known that restricted isometry property (RIP) constants (see Definition 3.2) can be used to prove exact recovery [15, 44]. Our first contribution is a novel connection between RIP constants and sharpness in the sense of (1.3). By adapting known proofs of exact recovery based on restricted isometry, we can describe the parameters (μ, r, ℓ, L) in terms of RIP constants. In particular, we show that there exist $\mu, r, \ell, L = \Theta(1)$ and $m = \Theta(n)$, so that w.h.p., the function $F_{r,\ell}$ is μ-sharp and L-Lipschitz. Almost identical results on dimension-independent conditioning for the other signal recovery problems are also presented in Theorem 4.1.
- *Optimization error and error bound*: In practice, we are unable to solve (1.1) exactly and must resort to numerical optimization. Using μ-sharpness, i.e., (1.3), we have that any ϵ-suboptimal solution[4] X to (1.1) or (1.2) satisfies $\left\|X - X^\natural\right\|_1 \leq \epsilon/\mu$. Using the estimate $\mu = \Theta(1)$, we are able to control how optimization error affects the distance to solution, $\left\|X - X^\natural\right\|_1$, a form of recovery error. Thus, μ-sharpness can be viewed as a particular form of an error bound, see (1.5) for a general definition. Compared to approaches [7, 22, 24, 51] based on Hoffman's eror bound [33] or other regularity conditions, e.g., strict complementarity [1], our approach based on the RIP yields a much clearer relationship between the parameters of the error bound, e.g., μ, r, ℓ, and the underlying dimension n. See the paragraph containing (1.5) in Section 1.1 for detailed discussions.
- *Measurement error and sensitivity of solutions*: Suppose that instead of observing $b = \mathcal{A}(X^\natural)$, we observe $\tilde{b} = \mathcal{A}(X^\natural) + \delta$ with a (possibly adversarial) noise term. It is known that (1.1) actually admits a unique *feasible* solution $X^\natural$ w.h.p. [10], thus (1.1) with b replaced by $\tilde{b}$ is likely to be infeasible. On the other hand, given any optimizer $\tilde{X}$ of (1.2) with $\tilde{b}$ in place of b, we may use sharpness in the noiseless case to bound $\left\|\tilde{X} - X^\natural\right\|_1 \leq O(\|\delta\|_1/\mu)$ in the noisy case (see Proposition 5.1).

  In a similar vein, we may consider a setting where b is corrupted by a sparse vector δ. Sharpness in the noiseless case tells us that the optimizer of (1.2) is *unchanged* if $|\mathrm{supp}(\delta)|/m = O(\mu)$, i.e., if up to an $O(\mu)$ fraction of the entries of b are corrupted (see Proposition 5.6 and Example 5.7 for a formal statement).[5]

  Thus, the parameter μ allows us to control how measurement error affects the recovery process. Our estimate $\mu = \Theta(1)$ shows that the optimizer is insensitive to noise in b.
- *Algorithms*: Since (1.1) has well-conditioned sharpness in the ℓ_1 norm, its sharpness in the standard ℓ_2 norm is necessarily ill-conditioned (see Proposition 6.2). This ill-conditioning causes both the theoretical performance guar-

[3] One may replace the ℓ_1 norm with other norms in the right hand side of these definitions (1.3) and (1.4). The benefit of the ℓ_1 norm, as we will discuss, is that the conditioning in our applications can be shown to be *dimension-independent*.

[4] Given an optimization problem $\min_{x\in\mathcal{X}} f(x)$ with minimum f^*, a feasible region $\mathcal{X}$ in some finite dimensional Euclidean space, and an objective function f, a vector x is an ϵ-suboptimal solution to $\min_{x\in\mathcal{X}} f(x)$ if $f(x) - f^* \leq \epsilon$ and $x \in \mathcal{X}$.

[5] Naturally, the allowed fraction of corrupted entries will satisfy $O(\mu) < 1/2$.

antees and the numeric performance of standard sharpness-aware first-order methods, which are based on the ℓ_2 norm, to deteriorate as the dimension increases (see Figure 1a). Thus, to solve large-scale instances of (1.1), we develop a new restarted mirror descent algorithm (RMD), which can be applied to *general* convex functions that are μ-sharp and L-Lipschitz in terms of an ℓ_p norm with $p \in [1, 2]$ (see Section 6). We may apply this algorithm to (1.2) with $p = 1$ to produce an ϵ-suboptimal solution in

$$O\left(\kappa^2 \log(n) \log\left(\frac{1}{\epsilon}\right)\right)$$

iterations. RMD's *logarithmic* dependence on n contrasts with the *linear* dependence on n that an ℓ_2-based algorithm would incur. Furthermore, each iteration of RMD can be implemented in time comparable to a single iteration of subgradient descent on (1.2).[6] These convergence guarantees also hold after appropriate modifications in the presence of corruption or noise (see Propositions 5.1 and 5.6) with guarantees depending on κ.

Our results on optimization error, measurement error, and algorithms above depend solely on the conditions (1.3) and (1.4) and are stated in terms of paramters (μ, r, ℓ, L). For a general problem, it may be that these conditions do not hold for *any* choice of these parameters. However, we will show that under standard statistical assumptions, not only do (1.3) and (1.4) hold for some choice of (μ, r, ℓ, L), but that these parameters are well-controlled. We emphasize again that the above results continue to hold similarly for other modern signal recovery problems.

1.1. Related work. To better position our work in the literature, in this section, we discuss related work on error bounds and conditioning, sharpness and conditioning in statistical recovery, and first-order methods for minimizing sharp functions.

Establishing error bounds and conditioning. The sharpness condition in (1.3) can be viewed as an error bound for (1.1). Error bounds and conditioning are central to optimization problems both structurally and algorithmically [1, 24, 34, 35, 38, 55]. Consider the following general error bound for (1.1):

$$\text{(1.5)}\quad \operatorname{tr}(X) + r\,\|\mathcal{A}(X) - b\| + \ell \operatorname{dist}(X, \mathbb{S}^n_+) - \operatorname{tr}(X^\natural) \geq \mu \left\|X - X^\natural\right\|^s, \text{ for all } X \in \mathbb{S}^n,$$

where, in addition to the r and ℓ introduced in (1.3), we also have the power s, and general norms throughout. Most methods for establishing error bounds either require that (1.1) (or the more general problem (P) introduced in later sections) is linear or its subdifferential graph is linear [7, 24, 33], or that regularity conditions such as Slater's condition [5, 54], strict complementarity [22, 24, 51], or nondegeneracy conditions [37] hold. Additional work has studied error bounds in generic semialgebraic settings [23]. Results of this type generally place an emphasis on establishing the power s (usually to 1 or 2) and provide only weak bounds on μ, r, and ℓ, that potentially depend on the dimension polynomially or exponentially. This is natural as the power s determines the linear or sublinear convergence rate [2, 34]. On the other hand, if μ, r, or ℓ depend polynomially (or even exponentially) on the dimension, then first-order algorithms may still be doomed to dimension-dependent convergence rates. This work connects the quantities μ, r, and ℓ to the quantitatively better-understood restricted isometry

[6] In the phase retrieval case, RMD requires performing two eigenvalue value decompositions per iteration, whereas subgradient descent requires performing one eigenvalue decomposition per iteration.

property (RIP). This connection gives precise estimates of the conditioning (μ, r, ℓ, L) in our settings and shows that these quantities are in fact dimension independent.

Sharpness and well conditioning in statistical recovery. A series of works [18, 19, 26, 36] studied the local landscape of *nonconvex* nonsmooth formulations in statistical recovery near the signal, particularly in low-rank matrix recovery. They showed that near a factored form of the ground truth $X^\natural$, the natural loss function is sharp, and the local condition number (in terms of the Frobenius or Schatten-2 norm) does not depend on the dimension directly. Hence, they can apply off-the-shelf algorithms, such as subgradient and prox-linear methods [21, 25] to achieve quick convergence to a factored form of $X^\natural$. Note that results of this type are *local*, and the algorithms require specific model knowledge to construct an appropriate initializer to achieve provable guarantees. In contrast, the methods developed here are convex optimization methods so they are not as sensitive to poor initialization.

First-order methods for minimizing sharp functions. Early work in this area established linear convergence of the subgradient method (with appropriate step sizes) for μ-sharp L-Lipschitz functions in the ℓ_2 norm [28, 32, 42]. These methods and their proofs are adapted to the ℓ_2 norm and incur a *linear dependence on the dimension* when applied to μ-sharp L-Lipschitz functions in the ℓ_1 norm.

Similar guarantees can be derived as a consequence of restarting schemes. Perhaps the earliest work discussing restarting schemes is [39], where a restarting scheme was developed for convex Hölder-smooth minimization in ℓ_p spaces. More recent work [38, 47, 48, 53] has used restarting schemes to accelerate variants of (sub)gradient descent in the presence of different growth conditions, e.g., a local guarantee of the form[7]

$$f(x) - \min_x f(x) \geq \mu \operatorname{dist}(x, \mathcal{X})^s$$

for all x close enough to the set of minimizers, $\mathcal{X}$ of a function f. Here, $\operatorname{dist}(\cdot, \cdot)$ is typically measured in the ℓ_2 norm and $s \in [1, \infty)$ captures the Hölderian growth of f. Taking $s = 1$ recovers the usual definition of sharpness. In this setting, restarted subgradient descent achieves a linear convergence rate of $O\left(\kappa^2 \log(1/\epsilon)\right)$ [53] where the condition number κ is measured in the ℓ_2 norm. Adaptive versions of these restart schemes have also been developed [45, 48] that do not require the sharpness parameter μ to be specified but incur an additional $\log(1/\epsilon)$ factor in the convergence rate.

We emphasize that much of the preceding literature on first-order methods for sharp convex functions focuses on the ℓ_2 setting. The only exception we know of is the work of [48], which develops a restarted mirror descent method for functions satisfying a property inspired by sharpness in the ℓ_2 setting. Roulet and d'Aspremont [48] show that their method can be used to achieve $O(\kappa^2 \log(n) \log(1/\epsilon))$ convergence on this particular function class. Unfortunately, in the ℓ_p setting with $p \in [1, 2)$, their proxy for sharpness is quite restrictive and does not allow solutions with sparseness or low-rank as is common in signal recovery. Our work in section 6 is closely related to [48] but uses a different specification of the restarting mechanism. This change allows us to use the natural definition of sharpness in an ℓ_p norm for $p \in [1, 2]$ (see Remark 6.7 for a thorough comparison).

1.2. Outline and notation.

Outline. We begin, in Section 2, by setting up notation and defining sharpness in the ℓ_1 norm for an abstract signal recovery problem. We will quantify sharpness

[7] Other forms of the necessary "growth" are possible in different settings.

and conditioning by a set of parameters (μ, r, ℓ, L). In Sections 3 and 4, we establish conditioning and sharpness under the Gaussian sensing models for the problems of sparse recovery, low-rank matrix and bilinear sensing, and covariance estimation in the absence of noise. Our proof relies on the well-known restricted isometry property (RIP). In section 5, we investigate how sharp instances of the abstract signal recovery problem behave in the presence of noise. In section 6, we describe and analyze an algorithm restarted mirror descent (RMD) that has linear and nearly dimension-independent convergence rates when applied to sharp functions in any ℓ_p or Schatten-p norms for $p \in [1, 2]$ (see Assumption 6.1). In conjunction with the results of Section 5, we may apply RMD to obtain linear nearly dimension-independent convergence rates even in the presence of noise for these statistical recovery problems.

A note on ℓ_p and Schatten-p norms. Let $p \in [1, \infty]$. We will overload notation and use $\|x\|_p$ for the standard ℓ_p norm when $x \in \mathbb{R}^n$ and the Schatten-p norm when $x \in \mathbb{R}^{n\times N}$. Recall that the nuclear norm is equivalent to the Schatten-1 norm, the Frobenius norm is equivalent to the Schatten-2 norm, and the spectral norm is equivalent to the Schatten-∞ norm. To streamline the text, we will also take "ℓ_p norm" of a matrix to mean its Schatten-p norm.

ℓ_0 "norm". Finally, we abuse notation and write $\|x\|_0$ for the size of the support of a vector or the rank of a matrix. We caution that this is *not* a norm.

2. Preliminaries.

In this section, we describe an abstract signal recovery problem and the convex optimization approach. Then, we define our measures of sharpness and conditioning.

2.1. Signal recovery via convex optimization.

Let V be the space of the signal, i.e., $V = \mathbb{R}^n$, $\mathbb{R}^{n\times N}$, or $\mathbb{S}^n$. We will equip V with the ℓ_1 norm. Again, "ℓ_1" means the standard ℓ_1 norm if $V = \mathbb{R}^n$ and the Schatten-1 norm if $V = \mathbb{R}^{n\times N}$ or $V = \mathbb{S}^n$. Let $\mathbb{R}^m$ be the space of the measurements equipped with the ℓ_p norm for some $p \in [1, 2]$ to be determined.

In the abstract signal recovery problem, we aim to recover an unknown element $x^\natural \in V$, called the signal, from a linear sensing operator, $\mathcal{A} : V \to \mathbb{R}^m$, and a vector of measurements, $b = \mathcal{A}(x^\natural) \in \mathbb{R}^m$. The convex optimization approach solves the following generalization of (1.1),

$$\text{(P)} \qquad \min_{x\in V} \left\{ f(x) : \begin{array}{l} \mathcal{A}(x) = b \\ x \in \mathcal{K} \end{array} \right\},$$

to recover $x^\natural$. Here, the function $f : V \to \mathbb{R}$ is convex and $\mathcal{K} \subseteq V$ is a closed convex cone (possibly all of V). They reflect our prior knowledge of $x^\natural$, e.g., if $x^\natural$ is sparse, we set $f(x) = \|x\|_1$.

Next, we define the sharpness and conditioning of (P).

2.2. Definition of sharpness and conditioning.

Given $x \in V$, and a closed nonempty set $\mathcal{X} \subseteq V$, let

$$\operatorname{dist}_1(x, \mathcal{X}) \coloneqq \min_{\bar{x}\in\mathcal{X}} \|\bar{x} - x\|_1 .$$

The following definition specializes the notion of sharpness for a convex function (also referred to as weak sharpness in the literature [30]) to the ℓ_1 setting.

DEFINITION 2.1 (Sharpness). *Let $\phi : V \to \mathbb{R}$ be a convex function, $\mathcal{X} \subseteq V$ a nonempty convex set, and let $\mu > 0$. We say that ϕ is μ-sharp around $\mathcal{X}$ w.r.t. the ℓ_1*

norm if $\mathcal{X} = \arg\min \phi$ *and*

$$\phi(\bar{x}) - \min_{x \in V} \phi(x) \geq \mu \operatorname{dist}_1(\bar{x}, \mathcal{X}), \qquad \forall \bar{x} \in V.$$

We extend the notion of sharpness to a problem of the form (P) as follows.

DEFINITION 2.2 (Sharpness of (P)). *Consider a problem of the form* (P). *Let* $x^\natural \in V$, $\mu > 0$, *and* $r, \ell \geq 0$. *We say that* (P) *is* (μ, r, ℓ) sharp around $x^\natural$ w.r.t. the ℓ_1 norm *if* $x^\natural$ *is feasible in* (P) *and*

$$F_{r,\ell}(x) \coloneqq f(x) + r \left\| \mathcal{A}(x) - b \right\|_p + \ell \operatorname{dist}_1(x, \mathcal{K}) \tag{2.1}$$

is μ*-sharp around* $x^\natural$. *Equivalently, if* $x^\natural$ *is feasible in* (P) *and*

$$\left\| x - x^\natural \right\|_1 \leq \frac{1}{\mu} \left(f(x) - f(x^\natural) + r \left\| \mathcal{A}(x) - b \right\|_p + \ell \operatorname{dist}_1(x, \mathcal{K}) \right), \qquad \forall x \in V. \tag{2.2}$$

If $\mathcal{K} = V$, *then* ℓ *is inconsequential, so we will simply say* (P) *is* (μ, r) *sharp around* $x^\natural$ *w.r.t. the* ℓ_1 *norm.*

Remark 2.3. Suppose (P) is (μ, r, ℓ)-sharp around $x^\natural$ w.r.t. the ℓ_1 norm for some $\mu > 0$, $r, \ell \geq 0$. Note that $x^\natural$ is necessarily the unique optimal solution of (P). Moreover, we have an error bound: If we numerically solve (P) and produce $\tilde{x} \in V$, then we may bound the distance $\left\| \tilde{x} - x^\natural \right\|_1$ according to (2.2). This bound holds even if $f(\tilde{x}) < f(x^\natural)$, which could happen as $\tilde{x}$ is not necessarily feasible for (P).

Recall the standard definition of the Lipschitz constant of a function specialized to the ℓ_1 norm.

DEFINITION 2.4 (Lipschitz constant). *We say that a convex function* $\phi : V \to \mathbb{R}$ *is* L*-Lipschitz w.r.t. the* ℓ_1 *norm if*

$$|\phi(x) - \phi(y)| \leq L \left\| x - y \right\|_1, \qquad \forall x, y \in V.$$

We are now ready to define the conditioning of (P). Note that the following definition depends implicitly on p, which we treat as fixed.

DEFINITION 2.5 (Conditioning and condition number). *Consider a problem of the form* (P) *and suppose* $x^\natural \in V$ *is its unique optimizer. Suppose for some* $r, \ell \geq 0$ *that* $F_{r,\ell}$ *is* μ*-sharp around* $x^\natural$ *and* L*-Lipschitz w.r.t. the* ℓ_1 *norm, with* $\mu, L > 0$. *The* conditioning *of* (P) *is measured by* (μ, r, ℓ, L) *and the* condition number *is* $\kappa = \frac{L}{\mu}$.

It is possible for a problem of the form (P) to not have well-defined conditioning according to this definition. For example, if $x^\natural$ is *not* the unique minimizer of (P), then its conditioning is not well-defined. Nevertheless, we will show that under standard statistical assumptions, many interesting signal recovery problems have well-defined conditioning under this definition.

3. RIP-based sharpness and conditioning in signal recovery. This section shows that under bounds on restricted isometry (defined below), sharpness in the ℓ_1 norm holds for three archetypal modern signal recovery problems: sparse vector recovery, low rank matrix sensing, and covariance estimation. Phase retrieval can also be thought of as a special case of covariance estimation. The signal to be recovered, the optimization formulation in (P), and the choices of norms are the following.

DEFINITION 3.1. *We consider the following signal recovery problems that can be formulated as* (P) *and the corresponding choices for* V, f, *and* $\mathcal{K}$.

- *Sparse vector recovery: recover a k-sparse vector $x^\natural \in V = \mathbb{R}^n$. In the formulation* (P)*, the objective is $f(x) = \|x\|_1$ and the set $\mathcal{K} = \mathbb{R}^n$.*
- *Low rank matrix sensing: recover a rank k matrix $X^\natural \in V = \mathbb{R}^{n\times N}$. In the formulation* (P)*, the objective is $f(X) = \|X\|_1$ and the set $\mathcal{K} = \mathbb{R}^n$.*
- *Covariance estimation: recover a rank k PSD matrix $X^\natural \in V = \mathbb{S}^n$. In the formulation* (P)*, the objective is the trace $f(X) = \operatorname{tr}(X)$ and the set $\mathcal{K}$ is the set of PSD matrices $\mathbb{S}^n_+$.*

The norm, $\|\cdot\|_p$ on the measurement space $\mathbb{R}^m$ is, as yet, unspecified in any of the three problems above. We have some freedom in the choice of p as long as crucial bounds on the restricted isometry property (RIP) hold in the ℓ_p norm.

DEFINITION 3.2 (Restricted isometry property). *Let k' be a positive integer and $\mathcal{A} : V \to \mathbb{R}^m$ be a linear operator. We will let $\mathrm{RIP}^-_{k'}(\mathcal{A})$ and $\mathrm{RIP}^+_{k'}(\mathcal{A})$ denote any valid* uniform *lower bound and upper bound on the quantity $\|\mathcal{A}(x)\|_p / \|x\|_2$ as $x \in V$ ranges over all elements in V with $\|x\|_0 \le k'$.*

Our main result of this section is the following:

PROPOSITION 3.3. *Consider one of the three problems defined in Definition* 3.1*. Let $c = 1$ for sparse vector recovery and $c = 2$ for the other two cases. Let $k' > 0$ and $\epsilon > 0$ and suppose $\mathrm{RIP}^+_{k'}(\mathcal{A}) \ge 1$ and*

$$\sqrt{\frac{k'}{ck}}\left(\frac{\mathrm{RIP}^-_{ck+k'}(\mathcal{A})}{\mathrm{RIP}^+_{k'}(\mathcal{A})}\right) \ge 1+\epsilon.$$

Then, (P) *is $\left(\frac{\epsilon}{2+\epsilon}, \sqrt{k'}, 2\right)$ sharp around $x^\natural$ w.r.t. ℓ_1. Furthermore, the Lipschitz constant L w.r.t. ℓ_1 is no more than $3 + \sqrt{k'}\,\mathrm{RIP}^+_1(\mathcal{A})$.*

To put this result in context, we will see in Section 4 that the premise of Proposition 3.3 holds once the sample size m is a small multiple of the recovery threshold w.h.p. with $\epsilon = 2$ and $k' = O(k)$, and $\mathrm{RIP}^+_1(\mathcal{A}) = O(1)$. Hence, Problem (P) is $\left(\frac{1}{2}, O\left(\sqrt{k}\right), 2\right)$ sharp around $x^\natural$ with a Lipschitz constant no more than $O\left(\sqrt{k}\right)$ w.h.p.

The rest of the section consists of proofs of Proposition 3.3 for the three different settings. We start with the proof of sharpness of matrix sensing and discuss how this proof can be modified for the other two settings. We then bound the Lipschitz constant for all three settings.

3.1. Proof of sharpness in matrix sensing. This section proves Proposition 3.3 for the setting of low-rank matrix sensing and uses many elements of the proofs for exact recovery [15, 44].

Let $\Delta \in \mathbb{R}^{n\times N}$ be arbitrary. Our goal is to show that

$$\left\|X^\natural + \Delta\right\|_1 + \sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p - \left\|X^\natural\right\|_1 \ge \frac{\epsilon}{2+\epsilon}\left\|\Delta\right\|_1. \tag{3.1}$$

Change of basis. Suppose $U_1 \in \mathbb{R}^{n\times n}$ and $U_2 \in \mathbb{R}^{N\times N}$ are orthonormal changes of bases. Note that the inequality (3.1) is invariant upon replacing (X, Δ) with $(U_1 X U_2^\intercal, U_1 \Delta U_2^\intercal)$ and replacing $\mathcal{A}$ with $\tilde{\mathcal{A}} : \tilde{\Delta} \mapsto \mathcal{A}(U_1^\intercal \tilde{\Delta} U_2)$. That is:

(3.1) $\iff$

$$\left\|U_1 X^\natural U_2^\intercal + U_1\Delta U_2^\intercal\right\|_1 + \sqrt{k'}\left\|\tilde{\mathcal{A}}(U_1\Delta U_2^\intercal)\right\|_p - \left\|U_1 X^\natural U_2^\intercal\right\|_1 \ge \frac{\epsilon}{2+\epsilon}\left\|U_1\Delta U_2^\intercal\right\|_1.$$

Furthermore, $\mathrm{RIP}^-_{k'}(\mathcal{A}) = \mathrm{RIP}^-_{k'}(\tilde{\mathcal{A}})$ and $\mathrm{RIP}^+_{k'}(\mathcal{A}) = \mathrm{RIP}^+_{k'}(\tilde{\mathcal{A}})$ for any k'. Thus, without loss of generality, we may assume that $X^\natural$ and Δ have the following form:

$$X^\natural = \begin{pmatrix} X^\natural_{1,1} & 0_{k\times(N-k)} \\ 0_{(n-k)\times k} & 0_{(n-k)\times(N-k)} \end{pmatrix} \qquad \text{and} \qquad \Delta = \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & \Delta_{2,2} \end{pmatrix}$$

where $X^\natural_{1,1}$ is a nonnegative $k\times k$ diagonal matrix and $\Delta_{2,2} \in \mathbb{R}^{(n-k)\times(N-k)}$ is diagonal with diagonal entries that are nonnegative and nonincreasing in magnitude.

Partitioning of the difference Δ. Let $\delta_{k^\perp} \coloneqq \mathrm{diag}(\Delta_{2,2}) \in \mathbb{R}^{\min(n,N)-k}$. The rest of the proof is based on the idea in [15, 44]. We decompose $\delta_{k^\perp} = \sigma_1 + \cdots + \sigma_t$ where each $\sigma_i \in \mathbb{R}^{\min(n,N)-k}$. Specifically, let σ_1 extract the first k' coordinates of $\delta_{k^\perp}$ as its first k' coordinates and be zero for other coordinates. We then let each subsequent σ_i extract the next k' coordinates of $\delta_{k^\perp}$ as its $(i-1)k'+1$-th to ik'-th coordinates and be 0 for other coordinates. Finally, σ_t may have fewer than k' nonzero coordinates of $\delta_{k^\perp}$ and its first $(t-1)k'$ coordinates are all zeros. Let Σ_i denote the matrix of size $n\times N$ with $\mathrm{Diag}(\sigma_i)$ in its bottom-right $(n-k)\times(N-k)$ block and zeros elsewhere.

Lower bound on $\|\Delta_{2,2}\|_1$. Since $\sigma_1, \ldots, \sigma_t$ have disjoint supports, we can lower bound $\|\Delta_{2,2}\|_1 = \|\delta_{k^\perp}\|_1$ by

$$\|\Delta_{2,2}\|_1 = \|\delta_{k^\perp}\|_1 \geq \sum_{i=1}^{t-1} \|\sigma_i\|_1 = k' \sum_{i=1}^{t-1} \frac{\|\sigma_i\|_1}{k'}.$$

Using $\delta_{k^\perp}$ is nonincreasing and each σ_i is k' sparse, we further have

$$\|\Delta_{2,2}\|_1 \geq k' \sum_{i=2}^{t} \|\sigma_i\|_\infty \geq \sqrt{k'} \sum_{i=2}^{t} \|\sigma_i\|_2 .$$

Next, we use the RIP to obtain

$$\begin{aligned} \|\Delta_{2,2}\|_1 &\geq \frac{\sqrt{k'}}{\mathrm{RIP}^+_{k'}} \sum_{i=2}^{t} \|\mathcal{A}(\Sigma_i)\|_p \\ &\overset{(a)}{\geq} \frac{\sqrt{k'}}{\mathrm{RIP}^+_{k'}} \left(\left\| \mathcal{A}\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,2} & 0 \end{pmatrix} + \mathcal{A}(\Sigma_1) \right\|_p - \|\mathcal{A}(\Delta)\|_p \right), \end{aligned}$$

where (a) is due to triangle inequality. Using the RIP condition again and $\mathrm{RIP}^+_{k'} \geq 1$, we have

$$\begin{aligned} \|\Delta_{2,2}\|_1 &\geq \sqrt{k'} \frac{\mathrm{RIP}^-_{2k+k'}}{\mathrm{RIP}^+_{k'}} \left\| \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0 \end{pmatrix} + \Sigma_1 \right\|_2 - \sqrt{k'} \|\mathcal{A}(\Delta)\|_p \\ &\geq \sqrt{k'} \frac{\mathrm{RIP}^-_{2k+k'}}{\mathrm{RIP}^+_{k'}} \left\| \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0 \end{pmatrix} \right\|_2 - \sqrt{k'} \|\mathcal{A}(\Delta)\|_p . \end{aligned}$$

Lastly, since $\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0 \end{pmatrix}$ has rank no more than $2k$, we have

$$\begin{aligned} \|\Delta_{2,2}\|_1 &\geq \sqrt{\frac{k'}{2k}} \frac{\mathrm{RIP}^-_{2k+k'}}{\mathrm{RIP}^+_{k'}} \left\| \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0 \end{pmatrix} \right\|_1 - \sqrt{k'} \|\mathcal{A}(\Delta)\|_p \\ &\geq (1+\epsilon) \left\| \begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0 \end{pmatrix} \right\|_1 - \sqrt{k'} \|\mathcal{A}(\Delta)\|_p . \end{aligned}$$

Putting things together. We are now ready to prove sharpness:

$$
\begin{aligned}
&\left\|X^\natural + \Delta\right\|_1 + \sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p - \left\|X^\natural\right\|_1 \\
\geq &\left\|\begin{pmatrix} X^\natural_{1,1} & 0 \\ 0 & \Delta_{2,2}\end{pmatrix}\right\|_1 - \left\|X^\natural\right\|_1 - \left\|\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0\end{pmatrix}\right\|_1 \\
&+ \sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p \\
\geq &\left\|\delta_{k^\perp}\right\|_1 - \left\|\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0\end{pmatrix}\right\|_1 + \sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p \qquad \text{(disjoint supports)} \\
= &\left(\frac{2}{2+\epsilon} + \frac{\epsilon}{2+\epsilon}\right)\left\|\delta_{k^\perp}\right\|_1 - \left\|\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0\end{pmatrix}\right\|_1 \\
&+ \sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p \\
\geq &\left(\frac{2}{2+\epsilon}(1+\epsilon) - 1\right)\left\|\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0\end{pmatrix}\right\|_1 + \frac{\epsilon}{2+\epsilon}\left\|\delta_{k^\perp}\right\|_1 \quad \text{(bound on } \left\|\delta_{k^\perp}\right\|_1\text{)} \\
&+ \left(1 - \frac{2}{2+\epsilon}\right)\sqrt{k'}\left\|\mathcal{A}(\Delta)\right\|_p \\
\geq &\frac{\epsilon}{2+\epsilon}\left\|\Delta\right\|_1 .
\end{aligned}
$$

This proves the claim (3.1) as $\Delta \in \mathbb{R}^{n\times N}$ was arbitrary.

3.2. Sharpness in sparse vector recovery. One can follow the proof for matrix sensing almost verbatim by treating the signal $x^\natural$ and the sensing vectors in $\mathcal{A}$ as diagonal matrices. The change-of-bases matrices may be picked as any permutation matrix $U_1 = U_2$. The change from $c = 2$ to $c = 1$ is justified by noting there is no extra off-diagonal term for the matrices considered in the argument. Thus, the matrix $\begin{pmatrix} \Delta_{1,1} & \Delta_{1,2} \\ \Delta_{2,1} & 0\end{pmatrix} = \begin{pmatrix} \Delta_{1,1} & \\ & 0\end{pmatrix}$ has rank bounded by k instead of $2k$.

3.3. Sharpness in covariance estimation. We may repeat the proof of Subsection 3.1 verbatim after replacing $V = \mathbb{R}^{n\times N}$ by $V = \mathbb{S}^n$. The change-of-bases matrices may be picked as any orthonormal matrix $U_1 = U_2$. We deduce that, under the assumptions of this proposition,

$$\left\|X\right\|_1 + \sqrt{k'}\left\|\mathcal{A}(X) - b\right\|_p$$

is $\frac{\epsilon}{2+\epsilon}$ sharp around $X^\natural$. Next, note that $\operatorname{tr}(X) + 2\operatorname{dist}_1(X, \mathbb{S}^n_+) \geq \left\|X\right\|_1$. We conclude that

$$\operatorname{tr}(X) + 2\operatorname{dist}_1(X, \mathbb{S}^n_+) + \sqrt{k'}\left\|\mathcal{A}(X) - b\right\|_p$$

is $\frac{\epsilon}{2+\epsilon}$ sharp around $X^\natural$.

3.4. Proof of Lipschitz continuity. We need to show that

$$f(x) + \sqrt{k'}\left\|\mathcal{A}(x) - b\right\|_p + 2\operatorname{dist}_1(x, K) \tag{3.2}$$

is $3 + \sqrt{k'}\operatorname{RIP}^+_1(\mathcal{A})$ Lipschitz with respect to the ℓ_1 norm.

By the triangle inequality, the objective functions $f(x) = \left\|x\right\|_1$ and $f(X) = \operatorname{tr}(X)$ are both 1-Lipschitz in terms of the ℓ_1 norm. In the covariance estimation setting, the distance function $\operatorname{dist}_1(X, \mathbb{S}^n_+)$ is also 1-Lipschitz in terms of the ℓ_1 norm by the triangle inequality.

Next, we show the function $\|\mathcal{A}(x)-b\|_p$ is RIP_1^+ Lipschitz. Indeed, for any $x_1, x_2 \in V$, let $\Delta = x_1 - x_2 = \sum_{i=1}^n \Delta_i$, where $\|\Delta_i\|_0 \le 1$. We have

$$\begin{aligned}
&\left| \|\mathcal{A}(x_1)-b\|_p - \|\mathcal{A}(x_2)-b\|_p \right| \\
&\le \|\mathcal{A}(x_1-x_2)\|_p = \left\| \mathcal{A}\left(\sum_{i=1}^n \Delta_i\right) \right\|_p \le \sum_{i=1}^n \|\mathcal{A}(\Delta_i)\|_p \\
&\overset{(a)}{\le} \sum_{i=1}^n \mathrm{RIP}_1^+ \|\Delta_i\|_2 \overset{(b)}{=} \sum_{i=1}^n \mathrm{RIP}_1^+ \|\Delta_i\|_1 = \mathrm{RIP}_1^+ \|\Delta\|_1 .
\end{aligned}$$

Here step (a) and (b) are due to the RIP and the fact that $\|\Delta_i\|_0 \le 1$ for all i.

Since $f(x)$ and $\mathrm{dist}_1(x,K)$ are 1 Lipschitz and $\|\mathcal{A}(x)-b\|_p$ is RIP_1^+ Lipschitz, our proof that the function in (3.2) is $\left(3+\sqrt{k'}\,\mathrm{RIP}_1^+(\mathcal{A})\right)$-Lipschitz with respect to the ℓ_1 norm is complete.

4. Conditioning, sensing models, and sample complexity. In this section, we describe different sensing models of $\mathcal{A}$ and show that once m exceeds certain thresholds, Problem (P) is well-conditioned in terms of ℓ_1 norm (for certain choices of the norm ℓ_p on the measurement space).

To prove this result, we first describe the precise sensing model in Subsection 4.1. Then we collect bounds on the RIP from the literature that ensure that the premise of Proposition 3.3 is satisfied. Due to an additional technicality in one of the sensing models (labeled "Covariance Estimation I" below), we give a separate proof of its conditioning in Subsection 4.2.

We will describe each sensing map $\mathcal{A}$ by specifying its adjoint map $\mathcal{A}^*$. More precisely, we equip $\mathbb{S}^n$ and $\mathbb{R}^{n\times N}$ with the trace inner product and equip $\mathbb{R}^n$ and $\mathbb{R}^m$ with the dot product. For each $i = 1, \dots, m$, let e_i denote the i-th standard basis of $\mathbb{R}^m$. We will specify the distribution of $\mathcal{A}^*(e_i) \in V$ and then recover $\mathcal{A} = (\mathcal{A}^*)^*$.

A summary of the sensing model, the thresholds, and the norms on $\mathbb{R}^m$ is described in Table 1. Our main result of this section is the following theorem.

THEOREM 4.1. *Suppose the norm on $\mathbb{R}^m$ and the sampling map $\mathcal{A}$ are described according to one of the scenarios in Table 1 and the signal $x^\natural$ satisfies $\|x\|_0 \le k$. If $m \ge CT(n,N,k)$, where $T(n,N,k)$ is defined in Table 1 and C is a numerical constant, then there are numerical constants $c_1, c_2 > 0$ such that w.h.p. the optimization problem* (P) *is $(\frac{1}{2}, \sqrt{c_1 k}, 2)$-sharp around $x^\natural$ and $F_{\sqrt{c_1 k},2}$ has a Lipschitz constant bounded by $\sqrt{c_2 k}$. Consequently,* (P) *has a condition number κ bounded by $2\sqrt{c_2 k}$.*

4.1. Sensing model, sample complexity, and proof via RIP. In this section, we first describe the sensing models and the thresholds $T(n,N,k)$ on m so that bounds on RIP hold for $\mathcal{A}$ when $\mathbb{R}^m$ is equipped with either the ℓ_1 or ℓ_2 norms. Then, we prove Theorem 4.1 by verifying the premise of Proposition 3.3.

Recall that $\mathrm{RIP}_{k'}^-(\mathcal{A})$ and $\mathrm{RIP}_{k'}^+(\mathcal{A})$ are any uniform lower and upper bound on $\|\mathcal{A}x\|_p / \|x\|_2$ as x ranges over elements of V with support or rank bounded by k'. They will be set to numerical constants c_1, c_2 below and could differ for different sensing models. The norm of $\mathbb{R}^m$ will be either $\ell_p = \ell_1$ or $\ell_p = \ell_2$.

Sparse vector recovery. For sparse vector recovery, the measurements are of the form

$$\mathcal{A}^*(e_i) = a_i, \quad i = 1, \dots, m,$$

Task	$\mathcal{A}^*(e_i)$	$T(n,N,k)$	ℓ_p norm on $\mathbb{R}^m$
Sparse vector recovery	$a_i \in \mathbb{R}^n$	$k\log(n/k)$	ℓ_1 or ℓ_2
Low rank matrix sensing I	$A_i \in \mathbb{R}^{n\times N}$	$\max\{n,N\}k$	ℓ_1 or ℓ_2
Low rank matrix sensing II	$a_i b_i^\top \in \mathbb{R}^{n\times N}$	$\max\{n,N\}k$	ℓ_1
Covariance estimation I	$a_i a_i^\top \in \mathbb{S}^n$	nk	ℓ_1
Covariance estimation II	$a_i a_i^\top - b_i b_i^\top \in \mathbb{S}^n$	nk	ℓ_1

Table 1: Description of different statistical signal recovery tasks and the thresholds for well-conditioning. The entries of A_i, a_i, and b_i are i.i.d. Gaussian random variables with appropriate scaling (see Section 4.1). The conditioning of (P) is measured by $(\frac{1}{2}, \sqrt{c_1 k}, 2, \sqrt{c_2 k})$ where $c_1, c_2 > 0$ are numerical constants.

where $a_i \overset{\text{i.i.d.}}{\sim} N(0, I/m)$ when $\ell_p = \ell_2$ and $a_i \overset{\text{i.i.d.}}{\sim} N\left(0, I/m^2\right)$ when $\ell_p = \ell_1$. In both settings, we may set w.h.p. $\mathrm{RIP}^+_{k'}(\mathcal{A}) = c_2$ and $\mathrm{RIP}^-_{k'}(\mathcal{A}) = c_1$ as long as $m \gtrsim k' \log(\frac{n}{k'})$,[8] where c_1 and c_2 are constants independent of k' satisfying $c_2/c_1 \le 1.1$ and $c_2 \ge 1$. This fact is proved in the ℓ_2 setting following [13, 27] using results on singular values of random matrices; a simple proof can be found in [4, Theorem 5.2]. This fact can be proved in the ℓ_1 setting using [41, Lemma 2.1] and [49, Lemma 4.4].[9]

Matrix sensing I. For this scenario of matrix sensing, the measurements are of the form

$$\mathcal{A}^*(e_i) = A_i, \quad i = 1, \dots, m,$$

where each measurement matrix $A_i \in \mathbb{R}^{n\times N}$ has Gaussian entries. Each entry of each matrix is sampled i.i.d. according to $N(0, 1/m)$ when $\ell_p = \ell_2$ setting, and according to $N\left(0, 1/m^2\right)$ when $\ell_p = \ell_1$. In both settings, we may set w.h.p. $\mathrm{RIP}^+_{k'}(\mathcal{A}) = c_2$ and $\mathrm{RIP}^-_{k'}(\mathcal{A}) = c_1$ as long as $m \gtrsim k' \max(n, N)$, where c_1 and c_2 are constants independent of k' satisfying $c_2/c_1 \le 1.1$ and $c_2 \ge 1$. This fact is proved in the ℓ_2 setting in [12, Theorem 2.3]. This fact is proved in the ℓ_1 setting in [36, Proposition 1].

Matrix sensing II. For this version of matrix sensing, which is more commonly known as bilinear sensing, the measurements are of the form

$$\mathcal{A}^*(e_i) = a_i b_i^\top, \quad i = 1, \dots, m,$$

and the measurement vectors $a_i \overset{\text{i.i.d.}}{\sim} N(0, I_n/m)$ and $b_i \overset{\text{i.i.d.}}{\sim} N(0, I_N/m)$. We equip $\mathbb{R}^m$ with the ℓ_1 norm. In this setting, we may set w.h.p. $\mathrm{RIP}^+_{k'}(\mathcal{A}) = c_2$ and $\mathrm{RIP}^-_{k'}(\mathcal{A}) = c_1$ as long as $m \gtrsim k' \max(n, N)$, where c_1 and c_2 and constants independent of k' satisfying $c_2/c_1 \le 4$ and $c_2 \ge 1$. This fact is proved according to [9, Theorem 2.2].

Covariance estimation I. For this scenario of covariance estimation, the measurements are of the form

$$\mathcal{A}^*(e_i) = a_i a_i^\top, \quad i = 1, \dots, m,$$

[8] We write $a \gtrsim b$ if $a \ge Cb$ for some numerical constant $C > 0$.

[9] Specifically, [41, Lemma 2.1] provides a RIP bound with a quantity called Guassian width. One then use [49, Lemma 4.4] to estimate the width and obtain the concrete RIP above.

where each measurement vector $a_i \overset{\text{i.i.d.}}{\sim} N(0, I_n/m)$. We equip $\mathbb{R}^m$ with the ℓ_1 norm. The proof of Theorem 4.1 for this setting differs from the proof of Theorem 4.1 for all other settings. This difference stems from the fact that $\langle \mathcal{A}^*(e_i), X^\natural \rangle$ does not have zero mean, thus biasing the output vector as discussed in [20, Section III.B]. Attempting to follow the same proof strategy will need $\mathrm{RIP}^+_{k'}(\mathcal{A})$ to be a numerical constant. However, the quantity $\mathrm{RIP}^+_{k'}(\mathcal{A})$ scales as $\sqrt{k'}$. This prevents us from applying Proposition 3.3 directly in this setting.

We provide a separate proof for Theorem 4.1(Covariance estimation I) in Subsection 4.2. The proof in this setting is completed by analyzing the conditioning of a related model of covariance estimation (Covariance estimation II).

Covariance estimation II. For this scenario of covariance estimation, the measurements are of the form

$$\mathcal{A}^*(e_i) = a_i a_i^\top - b_i b_i^\top, \quad i = 1, \dots, m,$$

where each measurement vector pair $a_i, b_i \overset{\text{i.i.d.}}{\sim} N(0, I_n/(2m))$. We equip $\mathbb{R}^m$ with the ℓ_1 norm. In this setting, we may set w.h.p. $\mathrm{RIP}^+_{k'}(\mathcal{A}) = c_2$ and $\mathrm{RIP}^-_{k'}(\mathcal{A}) = c_1$ as long as $m \gtrsim nk'$. Here, c_1 and c_2 are constants independent of k' satisfying $c_2/c_1 \le 4$ and $c_2 \ge 1$. To prove this fact, consider the operator $\bar{\mathcal{A}} : \mathbb{R}^{n\times n} \to \mathbb{R}^m$ such that $\bar{\mathcal{A}}^*(e_i) = (a_i - b_i)(a_i + b_i)^\top$. Note that $a_i - b_i$ and $a_i + b_i$ are independent of each other and $a_i - b_i, a_i + b_i \overset{\text{i.i.d.}}{\sim} N(0, I_n/m)$ for $i = 1, \dots, m$. Using [9, Theorem 2.2], we see that RIP holds for $\bar{\mathcal{A}}$ w.h.p. with $\mathrm{RIP}^+_{k'}(\bar{\mathcal{A}}) = c_2$ and $\mathrm{RIP}^-_{k'}(\bar{\mathcal{A}}) = c_1$ as long as $m \gtrsim nk'$. The proof is completed by noting that $[\bar{\mathcal{A}}(X)]_i = \langle (a_i - b_i)(a_i + b_i)^\top, X \rangle = [\mathcal{A}(X)]_i$ for any $X \in \mathbb{S}^n$.

The Gaussian assumption on the sensing vectors or matrices are assumed for simplicity. One can relax this condition to sub-Gaussian distributions as done in [4, Theorem 5.2], [41, Lemma 2.1], and [18, Theorem 6.4].

Proof of Theorem 4.1 for all settings except covariance estimation I. For all settings except covariance estimation I, as long as $m \gtrsim T(n, N, k')$, we may set w.h.p. $\mathrm{RIP}^+_{k'}(\mathcal{A}) = c_2$ and $\mathrm{RIP}^-_{k'}(\mathcal{A}) = c_1$. By setting $k' = \sqrt{C_1 k}$ for a large numerical constant C_1, we see the premise of Proposition 3.3 is satisfied and our proof is complete.

4.2. Conditioning of covariance estimation I. We provide a separate argument for Theorem 4.1 in the setting of Covariance estimation I.

In this setting, we have that $a_i \sim N(0, I_n/m)$ are i.i.d. We equip the space $\mathbb{R}^m$ with the ℓ_1 norm. Our goal is to show that (P) is sharp in terms of ℓ_1 norm with parameters $(\frac{1}{2}, \frac{1}{2}\sqrt{c_1 k}, 2)$ and Lipschitz continuous with Lipschitz constant $L = \sqrt{c_2 k}$ w.h.p. once $m \gtrsim nk$. We will do so by comparing Covariance estimation I with Covariance estimation II.

First, we replace the linear constraint in Problem (P) by $\mathcal{B}(X) = d$ where $\mathcal{B} : \mathbb{S}^n \to \mathbb{R}^{\lfloor m/2 \rfloor}$, and

$$\mathcal{B}^*(e_i) = \frac{1}{2} a_{2i-1} a_{2i-1}^\top - \frac{1}{2} a_{2i} a_{2i}^\top \quad d_i = \frac{1}{2} b_{2i-1} - \frac{1}{2} b_{2i}, \quad i = 1, \dots, \lfloor m/2 \rfloor. \tag{4.1}$$

This is distributed as an instance of Covariance estimation II. By Theorem 4.1 (Covariance estimation II), we know that once $m \gtrsim nk$, it holds that

$$\mathrm{tr}(X) + \sqrt{c_1 k}\, \|\mathcal{B}(X) - d\|_1 + 2\, \mathrm{dist}_1(X, \mathbb{S}^n_+) - \mathrm{tr}(X^\natural) \ge \frac{1}{2} \left\| X - X^\natural \right\|_1. \tag{4.2}$$

Note that for each $i = 1, \dots, \lfloor m/2 \rfloor$, by the construction (4.1), we have

$$|\langle \mathcal{B}^*(e_i), X\rangle - d_i| \le \frac{1}{2}|\langle \mathcal{A}^*(e_{2i-1}), X\rangle - b_{2i-1}| + \frac{1}{2}|\langle \mathcal{A}^*(e_{2i}), X\rangle - b_{2i}|.$$

Combining this fact with (4.2), we see that the function $\operatorname{tr}(X) + \frac{1}{2}\sqrt{c_1 k}\,\|\mathcal{A}(X) - b\|_1 + 2\operatorname{dist}_1(X, \mathbb{S}^n_+)$ is $\frac{1}{2}$ sharp around $X^\natural$ as well.

To prove the Lipschitz constant is bounded, we utilize [14, Lemma 3.1], which shows that w.h.p.

$$\|\mathcal{A}(X)\|_1 \le 1.1\,\|X\|_1.$$

Hence we see $\operatorname{tr}(X) + \frac{1}{2}\sqrt{c_1 k}\,\|\mathcal{A}(X) - b\|_1 + 2\operatorname{dist}_1(X, \mathbb{S}^n_+)$ is $3 + \sqrt{c_1 k}$ Lipschitz with respect to the ℓ_1 norm. This completes the proof of Theorem 4.1 (Covariance estimation I).

5. Sharp problem formulations in the presence of noise. In this section, we show that the sharpness of a problem (P) in the noiseless setting $b = \mathcal{A}(x^\natural)$ provides (algorithmically useful) information even in the noisy setting, where b is replaced by $\tilde{b} = \mathcal{A}(x^\natural) + \delta$ with δ small or sparse. We begin with the case where δ is small.

PROPOSITION 5.1. *Suppose* (P) *is* (μ, r, ℓ) *sharp around* $x^\natural$ *in the* ℓ_1 *norm. Let* $\delta \in \mathbb{R}^m$ *and set* $\tilde{b} = b + \delta$.

- *If* $\tilde{x}$ *minimizes*

$$\tilde{F}_{r,\ell}(x) := f(x) + r\left\|\mathcal{A}(x) - \tilde{b}\right\|_p + \ell \operatorname{dist}_1(x, \mathcal{K}),$$

  then $\left\|\tilde{x} - x^\natural\right\|_1 \le \frac{2r}{\mu}\|\delta\|_p$.
- *If* $\tilde{x}$ *minimizes*

$$\tilde{F}^{\text{thresh}}_{r,\ell}(x) := \max\left(\tilde{F}_{r,\ell}(x),\, F_{r,\ell}(x^\natural) + 3r\|\delta\|_p\right),$$

  then $\left\|\tilde{x} - x^\natural\right\|_1 \le \frac{4r}{\mu}\|\delta\|_p$. *Furthermore,* $\tilde{F}^{\text{thresh}}_{r,\ell}$ *is* $\frac{\mu}{2}$*-sharp around its optimizers.*

Remark 5.2. Suppose $F_{r,\ell}$ is L-Lipschitz with $f(x) = \|x\|_1$, $\mathcal{K} = V$ and $L \ge 1$ as in sparse vector recovery or low-rank matrix sensing. Then we have

$$L\left\|x^\natural\right\|_1 \ge \left|F_{r,\ell}(0) - F_{r,\ell}(x^\natural)\right| = \left|r\|b\|_p - \left\|x^\natural\right\|_1\right| \implies (L+1)\left\|x^\natural\right\|_1 \ge r\|b\|_p.$$

Hence, combining this inequality with the first item of Proposition 5.1, we have

$$\frac{\left\|\tilde{x} - x^\natural\right\|_1}{\|x^\natural\|_1} \le \frac{2L+2}{\mu}\frac{\|\delta\|_p}{\|b\|_p}.$$

Thus, we see that indeed the condition number of (P) controls the relative change of the solution to the relative perturbation to the data vector b:

$$\underbrace{\text{Relative change in solution}}_{\frac{\|\tilde{x}-x^\natural\|_1}{\|x^\natural\|_1}} \le \mathcal{O}(\kappa) \cdot \underbrace{\text{Relative change in data vector}}_{\frac{\|\delta\|_p}{\|b\|_p}}.$$

This is analogous to the use of condition number in linear equations.

Remark 5.3. When δ is nonzero, the penalization formulation $\tilde{F}_{r,\ell}$ is not necessarily a sharp function. However, the above proposition asserts that $\tilde{F}_{r,\ell}^{\text{thresh}}(x)$ is still sharp. Hence, we may hope to apply the methods described in Section 6, which apply to sharp functions: On the surface, evaluating the function $\tilde{F}_{r,\ell}^{\text{thresh}}(x)$ (and its subgradients) requires access to both $x^\natural$ and $\|\delta\|_p$. While we will not have access to these quantities in practice, that is of little consequence if we only plan to apply first-order methods (as we suggest in Section 6). Indeed, any first-order method applied to $\tilde{F}_{r,\ell}$ will behave equivalently to the first-order method applied to $\tilde{F}_{r,\ell}^{\text{thresh}}$ until an iterate $\tilde{x}$ satisfying $\left\|\tilde{x} - x^\natural\right\|_1 \leq \frac{4r}{\mu}\|\delta\|_p$ is found. In particular, the algorithms presented in Section 6 applied to $\tilde{F}_{r,\ell}$ will converge linearly to such a point with a rate depending on $\frac{\mu}{2}$. We emphasize that such a procedure is *adaptive* (to the noise level $\|\delta\|_p$) in both $\left\|\tilde{x} - x^\natural\right\|_1$ and the rate of convergence to $\tilde{x}$.

Proof of Proposition 5.1. For notational convenience, in this proof we will drop all subscripts r, ℓ.

By the triangle inequality, for all $x \in V$, we have

$$\left|F(x) - \tilde{F}(x)\right| = \left|r\|\mathcal{A}(x) - b\|_p - r\left\|\mathcal{A}(x) - \tilde{b}\right\|_p\right| \leq r\|\delta\|_p.$$

For the first claim, note that by optimality of $\tilde{x}$ in $\tilde{F}$, we have that

$$F(\tilde{x}) \leq \tilde{F}(\tilde{x}) + r\|\delta\|_p \leq \tilde{F}(x^\natural) + r\|\delta\|_p \leq F(x^\natural) + 2r\|\delta\|_p.$$

Combining this inequality with μ-sharpness of F around $x^\natural$ proves the first claim.

Consider the second claim. Note that $\tilde{F}(x^\natural) \leq F(x^\natural) + r\|\delta\|_p$ so that $\tilde{F}$ achieves values bounded above by the threshold value of $F(x^\natural) + 3r\|\delta\|_p$. Thus, the set of minimizers of $\tilde{F}^{\text{thresh}}(x)$ is given by $\mathcal{X} \coloneqq \left\{x \in V : \tilde{F}(x) \leq F(x^\natural) + 3r\|\delta\|_p\right\}$. Then, if $\tilde{x}$ minimizes $\tilde{F}^{\text{thresh}}(x)$, we must have

$$F(\tilde{x}) \leq \tilde{F}(\tilde{x}) + r\|\delta\|_p \leq F(x^\natural) + 4r\|\delta\|_p.$$

Combining this inequality with μ-sharpness of F around $x^\natural$ shows that $\left\|\tilde{x} - x^\natural\right\|_1 \leq \frac{4r}{\mu}\|\delta\|_p$.

It remains to show that $\tilde{F}^{\text{thresh}}(x)$ is $\mu/2$ sharp around its optimizers $\mathcal{X}$. By the definition of sharpness (see Definition 2.1), the goal is to show that for any $\bar{x} \in V \setminus \mathcal{X}$, there exists $\tilde{x} \in \mathcal{X}$ satisfying

$$\frac{\mu}{2}\|\bar{x} - \tilde{x}\|_1 \leq \tilde{F}^{\text{thresh}}(\bar{x}) - \tilde{F}^{\text{thresh}}(\tilde{x}).$$

Note that for any $\bar{x} \in V \setminus \mathcal{X}$ and $\tilde{x} \in \mathcal{X}$, we have $\tilde{F}^{\text{thresh}}(\bar{x}) = \tilde{F}(\bar{x})$ and $\tilde{F}^{\text{thresh}}(\tilde{x}) = F(x^\natural) + 3r\|\delta\|_p$.

Set $\tilde{x} = (1-\alpha)x^\natural + \alpha\bar{x}$ where

$$\alpha = \frac{2r\|\delta\|_p}{\tilde{F}(\bar{x}) - F(x^\natural) - r\|\delta\|_p}.$$

As $\tilde{F}(\bar{x}) > F(x^\natural) + 3r\|\delta\|_p$, we have that α is well-defined and $\alpha \in [0, 1]$. By convexity

of $\tilde{F}$,

$$\begin{aligned}
\tilde{F}(\tilde{x}) &\le (1-\alpha)\tilde{F}(x^\natural) + \alpha\tilde{F}(\bar{x}) \\
&\le (1-\alpha)(F(x^\natural) + r\left\|\delta\right\|_p) + \alpha\tilde{F}(\bar{x}) \\
&= F(x^\natural) + r\left\|\delta\right\|_p + \left(\frac{2r\left\|\delta\right\|_p}{\tilde{F}(\bar{x}) - F(x^\natural) - r\left\|\delta\right\|_p}\right)\left(\tilde{F}(\bar{x}) - F(x^\natural) - r\left\|\delta\right\|_p\right) \\
&= F(x^\natural) + 3r\left\|\delta\right\|_p.
\end{aligned}$$

We deduce that $\tilde{x} \in \mathcal{X}$.

Finally, we have

$$\begin{aligned}
\frac{\mu}{2}\left\|\bar{x} - \tilde{x}\right\|_1 &= (1-\alpha)\frac{\mu}{2}\left\|x^\natural - \bar{x}\right\|_1 \\
&\le \frac{1-\alpha}{2}\left(F(\bar{x}) - F(x^\natural)\right) \\
&\le \frac{1}{2}\left(\frac{\tilde{F}(\bar{x}) - F(x^\natural) - 3r\left\|\delta\right\|_p}{\tilde{F}(\bar{x}) - F(x^\natural) - r\left\|\delta\right\|_p}\right)\left(\tilde{F}(\bar{x}) - F(x^\natural) + r\left\|\delta\right\|_p\right) \\
&= \frac{1}{2}\left(\tilde{F}(\bar{x}) - F(x^\natural) - 3r\left\|\delta\right\|_p\right)\left(\frac{\tilde{F}(\bar{x}) - F(x^\natural) - 3r\left\|\delta\right\|_p + 4r\left\|\delta\right\|_p}{\tilde{F}(\bar{x}) - F(x^\natural) - 3r\left\|\delta\right\|_p + 2r\left\|\delta\right\|_p}\right) \\
&\le \tilde{F}(\bar{x}) - F(x^\natural) - 3r\left\|\delta\right\|_p.
\end{aligned}$$

Here, the second line follows by μ-sharpness of F, the third line follows by the definition of α, and the final line follows from the premise that $\tilde{F}(\bar{x}) > F(x^\natural) + 3r\left\|\delta\right\|_p$. □

Example 5.4. If the norm on the measurement space is $\ell_p = \ell_1$, then Proposition 5.1 is tight up to constants. Indeed, consider the sparse recovery problem below (parameterized by $\mu \in (0,1]$ and $r > 0$):

$$\begin{aligned}
&V = \mathbb{R}^2, \qquad x^\natural = e_1 \\
&\mathcal{A} = \frac{1}{r}\begin{bmatrix} 1 & 0 \\ \mu & 1-\mu \end{bmatrix}, \qquad b = \frac{1}{r}\begin{bmatrix} 1 \\ \mu \end{bmatrix}.
\end{aligned}$$

Note that

$$\begin{aligned}
F(x) &\coloneqq \left\|x\right\|_1 + r\left\|\mathcal{A}(x) - b\right\|_1 \\
&= |x_1| + |x_2| + |x_1 - 1| + |\mu(x_1 - 1) + (1-\mu)x_2| \\
&\ge 1 + \mu\left|x_1 - 1\right| + \mu\left|x_2\right| \\
&= F(x^\natural) + \mu\left\|x - x^\natural\right\|_1.
\end{aligned}$$

Thus, F is μ-sharp around $x^\natural$ w.r.t. the ℓ_1 norm. Now, suppose instead of b, we receive $\tilde{b} = b - \delta e_2$ where $\delta \in [0, \mu/r]$. A similar calculation shows that $\tilde{x} \coloneqq (1 - r\delta/\mu)e_1$ is the minimizer of $\tilde{F}(x)$:

$$\begin{aligned}
\tilde{F}(x) &\coloneqq \left\|x\right\|_1 + r\left\|\mathcal{A}(x) - \tilde{b}\right\|_1 \\
&= |x_1| + |x_2| + |x_1 - 1| + |\mu(x_1 - 1) + (1-\mu)x_2 + r\delta| \\
&\ge 1 + \mu\left|x_1 - 1 + \frac{r\delta}{\mu}\right| + \mu\left|x_2\right| \\
&= \tilde{F}(\tilde{x}) + \mu\left\|x - \tilde{x}\right\|_1.
\end{aligned}$$

We deduce that $\tilde{x}$, the minimizer of $\tilde{F}(x)$, satisfies $\left\|\tilde{x} - x^\natural\right\|_1 = \frac{r}{\mu}\|\delta\|_1$. This example can be extended to any $n, m \geq 2$ by padding.

Thus, if the only structural property of $(\mathcal{A}, b)$ that we are given is that $\mathcal{F}_{r,\ell}$ is sharp in the noiseless case and δ *may be chosen adversarially*, then a recovery error of $\left\|x - x^\natural\right\|_1 = O\left(\frac{r\|\delta\|_1}{\mu}\right)$ is optimal. This matches Proposition 5.1.

Remark 5.5. We believe the bound in Proposition 5.1 to be *suboptimal* if the error δ *is generated randomly.* For example, if we take the norm on the measurement space to be $\ell_p = \ell_2$ and generate $\mathcal{A}$ according to subsection 3.2, $\delta_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2/m)$, with m large enough, then Proposition 5.1 guarantees $\left\|\tilde{x} - x^\natural\right\|_1 = O\left(\sigma\sqrt{k}\right)$. One may compare this with the nearly statistically optimal error $\left\|\tilde{x} - x^\natural\right\|_2 = O\left(\sigma\sqrt{k\log(n)/m}\right)$ under the same assumptions [6]. It is not clear how to extend the nearly optimal result to the setting where the measurement space is equipped with the ℓ_p norm (for $p < 2$) and stochastic noise. Indeed, the argument in Proposition 5.1 considers only the norm on $\|\delta\|_p$ and is fundamentally unable to control cancellations in dealing with $\mathcal{A}(x) - \tilde{b}$. On the other hand, the argument in [6] specifically uses Euclidean properties of the ℓ_2 norm and does not extend to general ℓ_p norms. The choice $p = 1$ specifically is desirable as this enables robustness to grossly-but-sparsely-corrupted observations (see the following proposition). We leave the detailed analysis in the stochastic noise setting for future work.

The next proposition shows that if the measurement space is equipped with the $\ell_p = \ell_1$ norm and (P) is sharp in the noiseless case, then exact recovery continues to be possible (with linearly convergent algorithms) in the presence of grossly-but-sparsely-corrupted observations. Since the recovery procedure is invariant under permutations of the measurement space, we may assume without loss of generality that corruption occurs on the final portion of the measurement vector. Let $\mathbb{R}^{\tilde{m}}$ denote the measurement space. We decompose $\mathbb{R}^{\tilde{m}} = \mathbb{R}^m \times \mathbb{R}^{m'}$ so that the first m observations are uncorrupted and the remaining m' coordinates are arbitrarily corrupted.

PROPOSITION 5.6. *Fix V and let $\mathcal{A} : V \to \mathbb{R}^m$ and $\mathcal{B} : V \to \mathbb{R}^{m'}$ denote sensing operators. Equip the measurement space $\mathbb{R}^{\tilde{m}} = \mathbb{R}^m \times \mathbb{R}^{m'}$ with the $\ell_p = \ell_1$ norm. Suppose*

$$F_{r,\ell}(x) \coloneqq f(x) + r\left\|\mathcal{A}(x) - \mathcal{A}(x^\natural)\right\|_1 + \ell\,\mathrm{dist}_1(x, \mathcal{K})$$

is μ-sharp around $x^\natural$ w.r.t. the ℓ_1 norm. Let $\|\mathcal{B}\|_{1\to 1}$ denote the operator norm

$$\max_{z \in V:\, \|z\|_1 \leq 1} \|\mathcal{B}(z)\|_1 .$$

If $\|\mathcal{B}\|_{1\to 1} < \frac{\mu}{r}$, then the function

$$\tilde{F}_{r,\ell}(x) \coloneqq f(x) + r\left\|\begin{bmatrix}\mathcal{A}(x)\\ \mathcal{B}(x)\end{bmatrix} - \begin{bmatrix}\mathcal{A}(x^\natural)\\ \mathcal{B}(x^\natural) + \delta\end{bmatrix}\right\|_1 + \ell\,\mathrm{dist}_1(x, \mathcal{K}) \tag{5.1}$$

is $(\mu - r\|\mathcal{B}\|_{1\to 1})$-sharp around $x^\natural$.

Proof. Let $x \in V$. Then,

$$\begin{aligned}\tilde{F}_{r,\ell}(x) - \tilde{F}_{r,\ell}(x^\natural) &= F_{r,\ell}(x) - F_{r,\ell}(x^\natural) + r\left\|\mathcal{B}(x) - \mathcal{B}(x^\natural) - \delta\right\|_1 - r\|\delta\|_1\\ &\geq \mu\left\|x - x^\natural\right\|_1 - r\left\|\mathcal{B}(x) - \mathcal{B}(x^\natural)\right\|_1\\ &\geq (\mu - r\|\mathcal{B}\|_{1\to 1})\left\|x - x^\natural\right\|_1 .\end{aligned}$$

Here, the first inequality is the triangle inequality and the second inequality is the definition of the operator norm. ∎

Example 5.7. Consider the phase retrieval problem where an α-fraction of the observations are corrupted arbitrarily. Formally, consider the following procedure: Let $V = \mathbb{S}^n$, $m = \lceil (1-\alpha)\tilde{m} \rceil$ and $m' = \tilde{m} - m$. Fix $X^\natural \in \mathbb{S}^n_+ \subseteq V$ to be rank-one and let $\tilde{\mathcal{A}} : V \to \mathbb{R}^{\tilde{m}}$ be the random linear map

$$\tilde{\mathcal{A}}(X)_i = g_i^\intercal X g_i, \quad \text{where} \quad g_i \sim N(0, I/\tilde{m}).$$

Let $\delta \in \mathbb{R}^{\tilde{m}}$ denote an arbitrary vector, chosen possibly adversarially, with only the guarantee that $\operatorname{supp}(\delta) \subseteq [m+1, \tilde{m}]$. Set $\tilde{b} = \tilde{\mathcal{A}}(X^\natural) + \delta$.

Let $\mathcal{A} : V \to \mathbb{R}^m$ and $\mathcal{B} : V \to \mathbb{R}^{m'}$ denote the decomposition of $\tilde{\mathcal{A}} : V \to \mathbb{R}^{\tilde{m}}$ corresponding to the decomposition $\mathbb{R}^{\tilde{m}} = \mathbb{R}^m \times \mathbb{R}^{m'}$. Slightly abusing notation, we will also let δ denote the restriction of δ to $\mathbb{R}^{m'}$.

Our goal is to recover $X^\natural$ from

$$\tilde{F}_{r,\ell}(X) \coloneqq \operatorname{tr}(X) + r \left\| \tilde{\mathcal{A}}(X) - \tilde{b} \right\|_1 + \ell \operatorname{dist}_1(X, \mathbb{S}^n_+).$$

Now, suppose that

$$\min_{X \in \mathbb{S}^n_+} \left\{ \operatorname{tr}(X) : \begin{array}{l} \mathcal{A}(X) = \mathcal{A}(X^\natural) \\ X \in \mathbb{S}^n_+ \end{array} \right\}$$

is (μ, r, ℓ) sharp around $X^\natural$. Proposition 5.6 states that if $\|\mathcal{B}\|_{1\to 1} < \frac{\mu}{r}$, then $X^\natural$ is the unique minimizer of $\tilde{F}_{r,\ell}$ despite the corruption δ. By [52, Corollary 5.35], we know that w.h.p., $\|\mathcal{B}\|_{1\to 1} \le \left(\frac{\sqrt{\alpha\tilde{m}} + 2\sqrt{n}}{\sqrt{\tilde{m}}} \right)^2 \le 2\alpha + 8(n/\tilde{m})$. Combining these bounds, we have that $X^\natural$ is the unique minimizer of the sharp unconstrained minimization problem $\tilde{F}_{r,\ell}$ if $\alpha \lesssim \mu/r$ and $n/\tilde{m} \lesssim \mu/r$. In particular, μ/r controls the fraction of allowed gross corruption in the observation $\tilde{\mathcal{A}}(X^\natural)$.

6. First-order methods for non-Euclidean sharp minimization. We next turn to algorithms for minimizing the sharp nonsmooth formulations proposed in this paper. Our goal is to take advantage of the conditioning results (Theorem 4.1) and to design a first-order method tailored to sharp Lipschitz functions in the ℓ_1 norm.

Throughout this section, we make the following blanket assumption:

ASSUMPTION 6.1. *Let $V = \mathbb{R}^n$, $\mathbb{S}^n$, or $\mathbb{R}^{n\times N}$. In the third case, assume without loss of generality that $n \le N$. Assume that $f : V \to \mathbb{R}$ is a convex function with minimizers.*

Instead of directly designing algorithms for sharp Lipschitz functions w.r.t. the ℓ_1 norm, we will design algorithms for sharp Lipschitz functions w.r.t. the ℓ_p norm for all $p \in (1, 2]$ and view the $\ell_p = \ell_1$ case as a limiting case.

Let $p \in [1, 2]$. Recall that a function f is said to be μ-sharp w.r.t. the ℓ_p norm if

$$f(x) - f^* \ge \mu \operatorname{dist}_p \left(x, \arg\min_z f(z) \right) \qquad \forall x \in V,$$

where $f^* = \min_z f(z)$, and L-Lipschitz w.r.t. the ℓ_p norm if

$$|f(x) - f(y)| \le L \|x - y\|_p \qquad \forall x, y \in V.$$

We caution that the ℓ_p norm in this section is the norm on the signal space V and not the norm on the measurement space $\mathbb{R}^m$ as in prior sections. In prior sections, we always fixed $\ell_p = \ell_1$ in the signal space V.

Before we discuss our new algorithm, we first point out that sharp functions that are well-conditioned in ℓ_1 are necessarily poorly conditioned in ℓ_2.

PROPOSITION 6.2. *Suppose $f : V \to \mathbb{R}$ is μ-sharp and L-Lipschitz in the ℓ_1 norm around $x^\natural \in V$. If f is α-sharp and β-Lipschitz in the ℓ_2 norm around $x^\natural \in V$, then α and β must satisfy*

$$\alpha \leq L, \qquad \text{and} \qquad \beta \geq \mu\sqrt{n}.$$

Proof. First, suppose $V = \mathbb{R}^n$. Let $y = x^\natural + (\text{any 1-sparse vector})$. Then,

$$\alpha \left\|y - x^\natural\right\|_2 \leq f(y) - f(x^\natural) \leq L \left\|y - x^\natural\right\|_1.$$

We deduce that $\alpha \leq L$. Similarly, let $z = x^\natural + \mathbf{1}_n$ (the all-ones vector). Then,

$$\beta \left\|z - x^\natural\right\|_2 \geq f(z) - f(x^\natural) \geq \mu \left\|z - x^\natural\right\|_1.$$

We deduce that $\beta \geq \mu\sqrt{n}$.

The other two settings are proved analogously: take $y = x^\natural + (\text{any rank-1 matrix})$ and $z = x^\natural + I_n$ where I_n is the $n \times n$ identity matrix padded to $n \times N$. ∎

This proposition tells us that the functions $F_{r,\ell}$ from Theorem 4.1 have constant sharpness in the ℓ_2 norm and an $\Omega(\sqrt{n})$ Lipschitz constant in the ℓ_2 norm. Thus, first-order methods for sharp Lipschitz functions in the ℓ_2 norm [32, 42, 53] can at best guarantee a convergence rate of $O\left(n \log\left(\frac{1}{\epsilon}\right)\right)$. This dependence on n precludes the use of such algorithms (for example, the Polyak variant of the subgradient method `Polyak-Subgrad`) on large-scale instances of our signal recovery problems (see Figure 1a) and motivates our development of an algorithm with almost dimension-free iteration complexity in the next subsection.

6.1. Restarted Mirror Descent. We now describe the restarted mirror descent (RMD) algorithm, Algorithm 6.2. This algorithm generalizes similar algorithms for minimizing sharp functions in a Euclidean norm [32, 42, 53] and has nearly dimension-independent linear convergence rates that depend explicitly on sharpness (see Theorem 6.6) in an ℓ_p norm ($p \in [1, 2]$). Algorithm 6.2 can be applied to $F_{r,\ell}$, the sharp exact penalty formulation of (P), or any of its sharp perturbations (see Propositions 5.1 and 5.6).

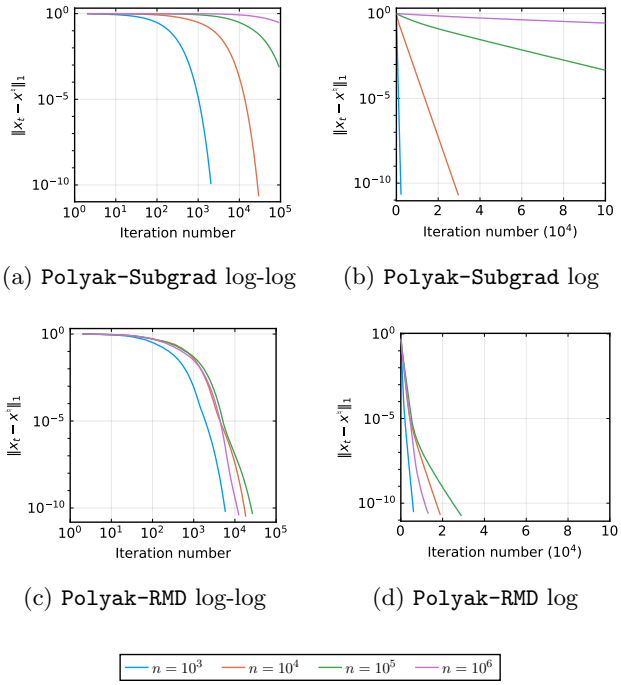
Recall the mirror descent algorithm and its guarantee [8, Theorem 4.2].

Algorithm 6.1 Mirror Descent

Given $f : V \to \mathbb{R}$, $\bar{x} \in V$, $\eta > 0$, $T \in \mathbb{N}$, $h : V \to \mathbb{R}$

- Let $x_0 = \bar{x}$, $\theta_0 = 0 \in V^*$
- For $t = 1, \ldots, T$
  - Let $g_t \in \partial f(x_{t-1})$ and set $\theta_t = \theta_{t-1} - \eta \cdot g_t$
  - Set $x_t = (\nabla h)^{-1}(\theta_t)$
- Output the x_t minimizing $f(x_t)$ among $t \in [0, T]$.

Here, V^* is the dual space of V. As a vector space, V^* can be identified with V via the standard inner product if $V = \mathbb{R}^n$ and via the trace inner product if $V = \mathbb{R}^{n \times N}$ or $V = \mathbb{S}^n$. The norm on V^* is the ℓ_q norm where q is the Hölder dual to p, i.e., the quantities q and p satisfies that $\frac{1}{p} + \frac{1}{q} = 1$.



(a) `Polyak-Subgrad` log-log (b) `Polyak-Subgrad` log

(c) `Polyak-RMD` log-log (d) `Polyak-RMD` log

Fig. 1: The convergence of `Polyak-Subgrad` and `Polyak-RMD` (described in Section 6.1) on sparse vector recovery for different values of n, the ambient dimension. The support size of $x^\natural$ is $k = 5$ and the sensing map $\mathcal{A}$ is generated as described in Section 4.1. The number of observations is $m = 2T$ where $T = (2k\log(n/k) + 1.25k + 1)$ is the current best estimate of the statistical threshold for sparse recovery [17]. See `https://github.com/alexlihengwang/sharpness_well_conditioning` for implementation details. Note `Polyak-Subgrad` deteriorates as the dimension n increases while `Polyak-RMD` is relatively insensitive to the dimension in terms of iteration number. Furthermore, one can see in subfigure (d) that `Polyak-RMD` achieves a linear convergence rate as predicted by Theorem 6.6 and Proposition 6.9.

LEMMA 6.3. *Let $h : V \to \mathbb{R}$ be differentiable and σ-strongly convex with respect to the ℓ_p norm. Suppose $f : V \to \mathbb{R}$ is convex and L-Lipschitz with respect to the ℓ_p norm. Then, the mirror descent algorithm initialized at $\bar{x}$, run for t iterations, with step-size η, produces $\tilde{x}$ such that*

$$f(\tilde{x}) - f(x^*) \leq \frac{L^2\eta}{2\sigma} + \frac{D_h(x^*||\bar{x})}{\eta t}$$

where $x^* \in V$ is any minimizer of f. Here, each iteration requires computing a single subgradient of f and applying $(\nabla h)^{-1}$ and arithmetic operations in V and V^*.

In the bound above, the quantity $D_h(\cdot||\cdot)$ is a Bregman divergence term. We elaborate on this term for our choice of the map h. Let $p \in (1, 2]$ and define $h_{\bar{x}}(x) := \frac{1}{2}\|x - \bar{x}\|_p^2$ for $\bar{x} \in V$. It is known [3, Theorem 1] that in each of the three setups in Assumption 6.1, that $h_{\bar{x}}$ is differentiable and $(p-1)$-strongly convex w.r.t. the ℓ_p norm. Furthermore, for all $x^* \in V$, the Bregman divergence (associated to $h_{\bar{x}}$) of x^*

with respect to $\bar{x}$ is

$$D_{h_{\bar{x}}}(x^*||\bar{x}) \coloneqq h_{\bar{x}}(x^*) - h_{\bar{x}}(\bar{x}) - \langle \nabla h_{\bar{x}}(\bar{x}), x^* - \bar{x} \rangle = h_{\bar{x}}(x^*) = \frac{1}{2} \|x^* - \bar{x}\|_p^2 .$$

Thus, if f is L-Lipschitz w.r.t. the ℓ_p norm, the output $\tilde{x} = \texttt{mirror}(h_{\bar{x}}, f, L, \bar{x}, t, \eta)$ has suboptimality bounded by

$$f(\tilde{x}) - f(x^*) \leq \frac{L^2 \eta}{2(p-1)} + \frac{\|x^* - \bar{x}\|_p^2}{2\eta T}.$$

This bound holds simultaneously for all minimizers $x^* \in V$ of f.

Remark 6.4. Consider a setup from Assumption 6.1 and let $p \in (1, 2]$. In each iteration of mirror descent (applied with $h_{\bar{x}}$), we must evaluate $(\nabla h_{\bar{x}})^{-1}$ on some input $\theta \in V^*$. By [46, Corollary 23.5.1], this is equivalent to evaluating $\nabla h_{\bar{x}}^*(\theta)$ where $h_{\bar{x}}^*$ is the convex conjugate of $h_{\bar{x}}$. Thus,

$$(\nabla h_{\bar{x}})^{-1}(\theta) = \nabla h_{\bar{x}}^*(\theta) = \bar{x} + \nabla \frac{1}{2} \|\theta\|_q^2 = \bar{x} + \frac{\operatorname{sign}(\theta) \circ |\theta|^{q-1}}{\|\theta\|_q^{q-2}}.$$

Here, q is the Hölder dual to p and the expression $\operatorname{sign}(\theta) \circ |\theta|^{q-1}$ is applied entrywise to the entries of θ if θ is a vector and to the singular values of θ if θ is a matrix.

Remark 6.5. In the matrix setting, computing the mirror map requires performing a singular value decomposition on θ. While this is an expensive operation, we emphasize that this is the same as the cost of simply *computing a subgradient* in all of the matrix recovery applications considered in this paper. For example, in the low-rank matrix sensing applications, computing the subgradient of

$$\|X\|_1 + r \|\mathcal{A}(X) - b\|_1$$

requires performing an SVD on X, and in covariance estimation, computing the subgradient of

$$\operatorname{tr}(X) + r \|\mathcal{A}(X) - b\|_1 + \ell \operatorname{dist}_1(X, \mathbb{S}_+^n)$$

requires performing an eigenvalue decomposition (EVD) on X, which can be obtained from an SVD of X or computed at a similar cost of an SVD of X. Thus, the work performed in each iteration of mirror descent is comparable to the work performed in each iteration of subgradient descent: two SVDs for mirror descent (one for the subgradient and one for the mirror map) and one SVD for subgradient descent. We expect that this full SVD may be replaced by a partial SVD near an optimal low-rank solution (following [31]) and leave this extension for future work.

Now consider the following restarted variant of mirror descent where the step size *and* mirror map h update at each restart.

Algorithm 6.2 $\texttt{RMD}(f, L, x_0, K, \{\eta_k\}, t, p)$

- For $k = 1, 2, \ldots, K$
  - Set $x_k \leftarrow \texttt{mirror}(h_{x_{k-1}}, f, L, x_{k-1}, t, \eta_k)$ where

$$h_{x_{k-1}}(x) \coloneqq \frac{1}{2} \|x - x_{k-1}\|_p^2$$

- Output x_K

Readers who prefer a more concrete description of RMD can find one for the setting of sparse recovery in Appendix A.

THEOREM 6.6. *Consider a setup from Assumption* 6.1 *and let* $p \in (1,2]$. *Suppose* $f : V \to \mathbb{R}$ *is* L*-Lipschitz and* μ*-sharp w.r.t. the* ℓ_p *norm and* $x_0 \in V$ *satisfies* $f(x_0) - f^* \le \epsilon_0$. *Let* $\epsilon_k = \epsilon_0 e^{-k/2}$, $K = \left\lceil 2\ln\left(\frac{\epsilon_0}{\epsilon}\right)\right\rceil$, $t = \left\lceil \frac{eL^2}{\mu^2(p-1)} \right\rceil$, *and* $\eta_k = \frac{(p-1)\epsilon_k}{L^2}$. *Then, each iterate* x_k *in* RMD$(f, L, x_0, K, \{\eta_k\}, t, p)$ *satisfies* $f(x_k) - f^* \le \epsilon_k$. *In particular, an* $0 < \epsilon \le \epsilon_0$*-optimizer can be computed in*

$$O\left(\frac{L^2}{\mu^2(p-1)} \log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$$

total mirror descent steps. If $f : V \to \mathbb{R}$ *is* L*-Lipschitz and* μ*-sharp w.r.t. the* ℓ_1 *norm, we may apply the above statement with* $p = 1 + \frac{1}{\ln n}$, *sharpness* μ, *and Lipschitz constant* eL *w.r.t. the* ℓ_p *norm.*

Proof. We first handle the case $p \in (1,2]$. The claim holds for $k = 0$. Now let $k \ge 1$. By μ-sharpness of f, there exists an optimizer x^* of f, satisfying $\mu \|x_{k-1} - x^*\|_p \le f(x_{k-1}) - f^*$. By Lemma 6.3, we have

$$\begin{aligned} f(x_k) - f^* &\le \frac{L^2\eta_k}{2(p-1)} + \frac{\|x^* - x_{k-1}\|_p^2}{2\eta_k t} \\ &\le \frac{L^2\eta_k}{2(p-1)} + \frac{e\epsilon_k^2}{2\mu^2\eta_k t} \\ &\le \frac{L^2\eta_k}{2(p-1)} + \frac{\epsilon_k^2(p-1)}{2\eta_k L^2} \\ &= \epsilon_k. \end{aligned}$$

The setting of L-Lipschitz, μ-sharp convex functions w.r.t. the ℓ_1 norm reduces to the setting of $p = 1 + \frac{1}{\ln(n)}$ by the bounds

$$\frac{1}{e}\|w\|_1 \le \|w\|_p \le \|w\|_1, \qquad \forall w \in V,$$

which hold [3] in each of the setups in Assumption 6.1. Specifically, these inequalities imply that f is μ-sharp and eL-Lipschitz w.r.t. the ℓ_p norm. ∎

Remark 6.7. In [48], the authors consider restarted versions of various accelerated first order methods for sharp problems including problems in non-Euclidean spaces. In the non-Euclidean setting, [48] suggests algorithms for functions f that satisfy the following proxy for sharpness:

$$f(x) - f(x^\natural) \ge \mu\sqrt{D_h(x^\natural \| x)}. \tag{6.1}$$

Here, $h : V \to \mathbb{R}$ is a *fixed* differentiable and σ-strongly convex function[10] and D_h is the Bregman divergence induced by h. This condition is inspired by sharpness in the ℓ_2 setting. Indeed, if we take $h(x) = \frac{1}{2}\|x\|_2^2$ to be the standard prox setup for a Euclidean space, then

$$\mu\sqrt{D_h(x^\natural \| x)} = \frac{\mu}{\sqrt{2}}\left\|x^\natural - x\right\|_2.$$

[10]The notation of [48] uses a 1-strongly convex function, but this only changes the value of μ by a factor of $\sqrt{\sigma}$.

In other words, under the standard Euclidean prox setup, (6.1) is equivalent to $\frac{\mu}{\sqrt{2}}$-sharpness w.r.t. the ℓ_2 norm. Assuming that the condition (6.1) holds, calculations almost identical to the proof of Theorem 6.6 show that restarted mirror descent with the mirror map h in every restart produces an ϵ-suboptimal solution in $O\left(\frac{L^2}{\mu^2\sigma}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$ mirror descent steps.

Unfortunately, (6.1) is *not* implied by sharpness in the ℓ_p norm if $p \in [1,2)$ and may be overly restrictive, as illustrated in Example 6.8 (see below). Specifically, in standard prox setups for mirror descent in ℓ_p norms or Schatten-p norms with $p \in [1,2)$, (6.1) cannot hold for any $\mu > 0$ unless $x^\natural$ has full support in the vector case or full rank in the matrix case.

Beyond the condition (6.1), [48] also considers extensions allowing for (possibly only local) Hölderian growth $f(x) - f(x^\natural) \geq \frac{\mu}{s}(D_h(x^\natural||x))^{s/2}$ for some $s \geq 1$. When f is nonsmooth and $s > 1$, [48, Corollary 5.4] guarantees an iteration complexity of $O(\epsilon^{-2\frac{s-1}{s}})$ after entering the local region. This guarantee can be improved to $O(\epsilon^{-\frac{s-1}{s}})$ for certain non-smooth problems by using smoothing (see [48, Proposition 5.5]). However, neither of these results offer an approach to achieve the $O(\log(1/\epsilon))$ iteration complexity achieved by `RMD`, unless $s = 1$.

Example 6.8. Let $V = \mathbb{R}^n$ and assume that $n \geq 2$. Let $p \in [1,2)$. Let $f(x) := \left\|x - x^\natural\right\|_p$ where $x^\natural = e_1$. Clearly, f is 1-sharp around $x^\natural$ and 1-Lipschitz w.r.t. the ℓ_p norm.

Two standard prox setups for mirror descent [40] in this setting are to take $h(x) = \frac{1}{2}\|x\|_p^2$ (in the unbounded case) or $h(x) = \frac{1}{p}\|x\|_p^p$ (in the bounded case); see [40, Theorem 2.1] for the corresponding strong convexity parameters. Consider the first setting, i.e., $h(x) = \frac{1}{2}\|x\|_p^2$. Letting $x_\epsilon = x^\natural + \epsilon e_2$, we have by Bernoulli's inequality that

$$\begin{aligned}\sqrt{D_h\left(x^\natural||x_\epsilon\right)} &= \sqrt{\frac{1}{2}(1+|\epsilon|^p)^{2/p} - \frac{1}{2} - \langle e_1, \epsilon e_2\rangle} \\ &= 2^{-1/2}\sqrt{(1+|\epsilon|^p)^{2/p} - 1} \\ &\geq p^{-1/2}\,|\epsilon|^{p/2}\,.\end{aligned}$$

Similarly, suppose $h(x) = \frac{1}{p}\|x\|_p^p$. Then,

$$\sqrt{D_h(x^\natural||x_\epsilon)} = \sqrt{\frac{1}{p}(1+|\epsilon|^p) - \frac{1}{p} - \langle e_1, \epsilon e_2\rangle} = p^{-1/2}\,|\epsilon|^{p/2}\,.$$

On the other hand, $f(x_\epsilon) - f(x^\natural) \leq |\epsilon|$. Thus, letting $|\epsilon| \to 0$, we see that

$$f(x_\epsilon) - f(x^\natural) \geq \mu\sqrt{D_h(x^\natural||x_\epsilon)}$$

cannot hold for *any* $\mu > 0$ in either of the standard prox setups.

An almost identical argument shows that a local Hölderian error bound of the form

$$f(x_\epsilon) - f(x^\natural) \geq \frac{\mu}{s}D_h(x^\natural||x_\epsilon)^{s/2} \qquad \forall \epsilon \text{ small}$$

can only hold with a positive $\mu > 0$ if $s \geq 2/p$. We conclude that the best guarantee one could achieve by applying[11] [48, Corollary 5.4] or [48, Proposition 5.5]

[11] This *assumes* that the conditions of [48, Corollary 5.4, Proposition 5.5] hold.

naïvely to the objective function $f(x) = \|x - \bar{x}\|_p$ would be an iteration complexity of $O\left(\epsilon^{(p-2)/2}\right)$. Heuristically substituting $p \leftarrow 1$ gives $O\left(\epsilon^{-1/2}\right)$ contrasting with the $O\left(\log(1/\epsilon)\right)$ iteration complexity achieved by `RMD`.

The following variant of Algorithm 6.2 replaces knowledge of μ with knowledge of f^*. This "Polyak" variant achieves the same convergence rate as Algorithm 6.2 with the optimal choice of μ.

Algorithm 6.3 `Polyak-RMD`$(f, L, x_0, K, \{\eta_k\}, f^*, \{\epsilon_k\}, p)$

- For $k = 1, 2, \ldots, K$
  - Run mirror descent `mirror`$(h_{x_{k-1}}, f, L, x_{k-1}, \infty, \eta_k)$ until it finds an iterate x_k satisfying
$$f(x_k) - f^* \leq \epsilon_k$$
- Output x_K

The convergence guarantee for Polyak-RMD (Algorithm 6.3) and its proof are entirely identical to that of Algorithm 6.2.

PROPOSITION 6.9. *Consider a setup from Assumption* 6.1 *and let* $p \in (1, 2]$. *Suppose* $f : V \to \mathbb{R}$ *is* L*-Lipschitz and* μ*-sharp w.r.t. the* ℓ_p *norm and* $x_0 \in V$ *satisfies* $f(x_0) - f^* \leq \epsilon_0$. *Let* $\epsilon_k = \epsilon_0 e^{-k/2}$, $K = \left\lceil 2 \ln\left(\frac{\epsilon_0}{\epsilon}\right)\right\rceil$, *and* $\eta_k = \frac{(p-1)\epsilon_k}{L^2}$. *Then, each iterate* x_k *in* `Polyak-RMD`$(f, L, x_0, K, \{\eta_k\}, f^*, \{\epsilon_k\}, p)$ *is computed in at most* $\left\lceil \frac{eL^2}{\mu^2(p-1)} \right\rceil$ *mirror descent steps. In particular, an* $0 < \epsilon \leq \epsilon_0$*-minimizer can be computed in*

$$O\left(\frac{L^2}{\mu^2(p-1)} \log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$$

total mirror descent steps. If $f : V \to \mathbb{R}$ *is* L*-Lipschitz and* μ*-sharp w.r.t. the* ℓ_1 *norm, we may apply the above statement with* $p = 1 + \frac{1}{\ln n}$, *sharpness* μ, *and Lipschitz constant* eL *w.r.t. the* ℓ_p *norm.*

Remark 6.10. The basic algorithmic structure of Algorithms 6.2 and 6.3 can also be extended to an adaptive variant. Specifically, it is possible to design an adaptive version of Algorithm 6.2 (similar in spirit to [45]) that does not require knowledge of either f^* or μ but would have a convergence rate of $O\left(\frac{L^2}{\mu^2(p-1)}\left(\log\frac{\epsilon_0}{\epsilon}\right)^2\right)$.

7. Discussion. This paper shows that for various statistical signal recovery tasks, once the sample size is greater than a constant multiple of the recovery threshold, then the convex optimization problem becomes well-conditioned in the sense of sharpness w.r.t. the ℓ_1 or Schatten-1 norm. In turn, this fact shows the optimization problem is robust to measurement error and optimization error. Furthermore, the newly developed algorithm `RMD` is able to achieve nearly-dimension-independent convergence rates.

We hope this paper induces interest in the interplay between statistics and convex optimization, especially in nonsmooth formulations and algorithms for these nonsmooth formulations. In particular, the following three directions might be of interest for future investigations:

- **Good conditioning beyond RIP**: This paper considers recovery tasks where strong RIP bounds have been established. Can we prove conditioning results for other statistical problems such as matrix completion [16] and phase

retrieval with coded diffraction patterns [11] where standard RIP bounds may not hold?

- **Adaptive penalty parameters**: Given a concrete statistical model, we can give upper bounds on r and ℓ so that the corresponding function $F_{r,\ell}$ is μ-sharp. On the other hand, determining these parameters may be difficult without a precise statistical model. Thus, are there algorithmic ways of estimating or adaptively choosing these two parameters?
- **Beyond convexity**: In our current setup, the objective function in (P) is required to be convex. For some applications in signal recovery, e.g. [29, 50], it is more convenient to use a nonconvex objective function. Can we extend the results of the current paper to those settings?

Appendix A. Concrete description of RMD for sparse recovery.

This appendix specializes RMD (Algorithm 6.2) to the sparse recovery setting, while fully unpacking the inner steps of Mirror Descent as well. This appendix does not present any new results beyond those already contained in the main text.

Our goal is to minimize

$$f(x) = \|x\|_1 + r\,\|Ax - b\|_1\,, \tag{A.1}$$

where $b = Ax^\natural$, over choices $x \in \mathbb{R}^n$. We will assume that $f(x)$ is μ-sharp and L-Lipschitz with respect to the ℓ_1 norm. Then, Theorem 6.6 guarantees that the following algorithm will output a $\tilde{x}$ satisfying $f(\tilde{x}) \le \epsilon$ after $O\left(\frac{L^2}{\mu^2}\ln(n)\ln\left(\frac{r\|b\|_1}{\epsilon}\right)\right)$ iterations. Using μ-sharpness one more time gives $\left\|x^\natural - \tilde{x}\right\|_1 \le \frac{\epsilon}{\mu}$.

Algorithm A.1 RMD-SR$(r, A, b, L, \epsilon, \mu)$

- Initialize $\bar{x}_0 = 0 \in \mathbb{R}^n$, $\epsilon_0 = r\,\|b\|_1$, and $\eta = \frac{\epsilon_0}{\ln(n)L^2}$
- Let $K = \left\lceil 2\ln\frac{\epsilon_0}{\epsilon}\right\rceil$, $t = \left\lceil \frac{eL^2\ln(n)}{\mu^2}\right\rceil$, $p = 1 + \frac{1}{\ln(n)}$, $q = 1 + \ln(n)$
- For $k = 1, 2, \ldots, K$
  - Initialize $x_0 = \bar{x}_{k-1}$, $\theta_0 = 0 \in \mathbb{R}^n$
  - Set $\eta \leftarrow \eta/\sqrt{e}$
  - For $i = 1, \ldots, t$
    - Let $g_i \in \partial f(x_{i-1})$
    - Let $\theta_i = \theta_{i-1} - \eta g_i$
    - Let $x_i = \bar{x}_{k-1} + \mathrm{sign}(\theta_i) \circ |\theta_i|^{q-1} \,/\, \|\theta_i\|_q^{q-2}$
  - Set $\bar{x}_k = \arg\min_{x\in\{x_0,\ldots,x_t\}} f(x)$
- Output $\bar{x}_K$

One can see that RMD-SR is a double-loop method, where the outer iterates $\bar{x}_0, \ldots, \bar{x}_K$ are each computed using an auxiliary sequence of iterates $x_0, \ldots, x_t$. The stepsize, η, that is used in the inner loop (Mirror Descent) decays geometrically with the outer iterations. Within each inner loop, the function that maps gradients θ_i to iterates x_i, gets "re-centered" around the most recent outer iterate $\bar{x}_{k-1}$. This map raises the magnitude of each entry of θ to a large power and then normalizes so that $\|x_i - \bar{x}_{k-1}\|_p = \|\theta\|_q$. This is exactly the behavior of RMD (Algorithm 6.2) in the general case as well.

Acknowledgments. Lijun Ding would like to thank Dmitriy Drusvyatskiy and Mateo Diaz for helpful discussions.

References.

[1] F. Alizadeh, J.-P. A. Haeberly, and M. L. Overton, *Complementarity and nondegeneracy in semidefinite programming*, Mathematical programming, 77 (1997), pp. 111–128.

[2] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, *Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality*, Mathematics of operations research, 35 (2010), pp. 438–457.

[3] K. Ball, E. A. Carlen, and E. H. Lieb, *Sharp uniform convexity and smoothness inequalities for trace norms*, Inequalities: Selecta of Elliott H. Lieb, (2002), pp. 171–190.

[4] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, *A simple proof of the restricted isometry property for random matrices*, Constructive approximation, 28 (2008), pp. 253–263.

[5] H. H. Bauschke, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, jameson's property (g), and error bounds in convex optimization*, Mathematical Programming, 86 (1999), pp. 135–160.

[6] A. Belloni, V. Chernozhukov, and L. Wang, *Square-root lasso: pivotal recovery of sparse signals via conic programming*, Biometrika, 98 (2011), pp. 791–806.

[7] J. Bolte, T. P. Nguyen, J. Peypouquet, and B. W. Suter, *From error bounds to the complexity of first-order descent methods for convex functions*, Mathematical Programming, 165 (2017), pp. 471–507.

[8] S. Bubeck et al., *Convex optimization: Algorithms and complexity*, Foundations and Trends® in Machine Learning, 8 (2015), pp. 231–357.

[9] T. T. Cai and A. Zhang, *Rop: Matrix recovery via rank-one projections*, The Annals of Statistics, 43 (2015), pp. 102–138.

[10] E. J. Candès and X. Li, *Solving quadratic equations via phaselift when there are about as many equations as unknowns*, Foundations of Computational Mathematics, 14 (2014), pp. 1017–1026.

[11] E. J. Candes, X. Li, and M. Soltanolkotabi, *Phase retrieval from coded diffraction patterns*, Applied and Computational Harmonic Analysis, 39 (2015), pp. 277–299.

[12] E. J. Candes and Y. Plan, *Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements*, IEEE Transactions on Information Theory, 57 (2011), pp. 2342–2359.

[13] E. J. Candes, J. K. Romberg, and T. Tao, *Stable signal recovery from incomplete and inaccurate measurements*, Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 59 (2006), pp. 1207–1223.

[14] E. J. Candès, T. Strohmer, and V. Voroninski, *Phaselift: Exact and stable signal recovery from magnitude measurements via convex programming*, Communications on Pure and Applied Mathematics, 66 (2013), pp. 1241–1274.

[15] E. J. Candes and T. Tao, *Decoding by linear programming*, IEEE transactions on information theory, 51 (2005), pp. 4203–4215.

[16] E. J. Candès and T. Tao, *The power of convex relaxation: Near-optimal matrix completion*, IEEE Transactions on Information Theory, 56 (2010), pp. 2053–2080.

[17] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, *The convex geometry of linear inverse problems*, Foundations of Computational

mathematics, 12 (2012), pp. 805–849.
[18] V. Charisopoulos, Y. Chen, D. Davis, M. Díaz, L. Ding, and D. Drusvyatskiy, *Low-rank matrix recovery with composite optimization: good conditioning and rapid convergence*, Foundations of Computational Mathematics, 21 (2021), pp. 1505–1593.
[19] V. Charisopoulos, D. Davis, M. Díaz, and D. Drusvyatskiy, *Composite optimization for robust blind deconvolution*, arXiv preprint arXiv:1901.01624, (2019).
[20] Y. Chen, Y. Chi, and A. J. Goldsmith, *Exact and stable covariance estimation from quadratic sampling via convex programming*, IEEE Transactions on Information Theory, 61 (2015), pp. 4034–4059.
[21] D. Davis, D. Drusvyatskiy, K. J. MacPhee, and C. Paquette, *Subgradient methods for sharp weakly convex functions*, Journal of Optimization Theory and Applications, 179 (2018), pp. 962–982.
[22] L. Ding and M. Udell, *A strict complementarity approach to error bound and sensitivity of solution of conic programs*, Optimization Letters, (2022), pp. 1–24.
[23] D. Drusvyatskiy, A. D. Ioffe, and A. S. Lewis, *Generic minimizing behavior in semialgebraic optimization*, SIAM Journal on Optimization, 26 (2016), pp. 513–534.
[24] D. Drusvyatskiy and A. S. Lewis, *Error bounds, quadratic growth, and linear convergence of proximal methods*, Mathematics of Operations Research, 43 (2018), pp. 919–948.
[25] D. Drusvyatskiy and C. Paquette, *Efficiency of minimizing compositions of convex functions and smooth maps*, Mathematical Programming, 178 (2019), pp. 503–558.
[26] J. C. Duchi and F. Ruan, *Solving (most) of a set of quadratic equalities: Composite optimization for robust phase retrieval*, Information and Inference: A Journal of the IMA, 8 (2019), pp. 471–529.
[27] A. Edelman, *Eigenvalues and condition numbers of random matrices*, SIAM journal on matrix analysis and applications, 9 (1988), pp. 543–560.
[28] I. I. Eremin, *The relaxation method of solving systems of inequalities with convex functions on the left-hand side*, in Doklady Akademii Nauk, vol. 160, Russian Academy of Sciences, 1965, pp. 994–996.
[29] J. Fan, L. Ding, Y. Chen, and M. Udell, *Factor group-sparse regularization for efficient low-rank matrix recovery*, Advances in neural information processing Systems, 32 (2019).
[30] M. C. Ferris, *Weak sharp minima and exact penalty functions*, tech. report, University of Wisconsin-Madison Department of Computer Sciences, 1988.
[31] D. Garber and A. Kaplan, *On the efficient implementation of the matrix exponentiated gradient algorithm for low-rank matrix optimization*, Mathematics of Operations Research, (2022).
[32] J.-L. Goffin, *On convergence rates of subgradient optimization methods*, Mathematical programming, 13 (1977), pp. 329–347.
[33] A. J. Hoffman, *On approximate solutions of systems of linear inequalities*, in Selected Papers Of Alan J Hoffman: With Commentary, World Scientific, 2003, pp. 174–176.
[34] P. R. Johnstone and P. Moulin, *Faster subgradient methods for functions with hölderian growth*, Mathematical Programming, 180 (2020), pp. 417–450.
[35] A. S. Lewis, *Nonsmooth optimization: conditioning, convergence and semialgebraic models*, in Proceedings of the International Congress of Mathematicians,

Seoul, vol. 4, 2014, pp. 872–895.
[36] X. Li, Z. Zhu, A. Man-Cho So, and R. Vidal, *Nonconvex robust low-rank matrix recovery*, SIAM Journal on Optimization, 30 (2020), pp. 660–686.
[37] M. V. Nayakkankuppam and M. L. Overton, *Conditioning of semidefinite programs.*, Mathematical programming, 85 (1999).
[38] I. Necoara, Y. Nesterov, and F. Glineur, *Linear convergence of first order methods for non-strongly convex optimization*, Mathematical Programming, 175 (2019), pp. 69–107.
[39] A. S. Nemirovski and Y. E. Nesterov, *Optimal methods of smooth convex minimization*, Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki, 25 (1985), pp. 356–369.
[40] Y. Nesterov and A. Nemirovski, *On first-order algorithms for l1/nuclear norm minimization*, Acta Numerica, 22 (2013), pp. 509–575.
[41] Y. Plan and R. Vershynin, *Dimension reduction by random hyperplane tessellations*, Discrete & Computational Geometry, 51 (2014), pp. 438–461.
[42] B. T. Polyak, *Minimization of unsmooth functionals*, USSR Computational Mathematics and Mathematical Physics, 9 (1969), pp. 14–29.
[43] S. Portnoy and R. Koenker, *The gaussian hare and the laplacian tortoise: computability of squared-error versus absolute-error estimators*, Statistical Science, 12 (1997), pp. 279–300.
[44] B. Recht, M. Fazel, and P. A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, SIAM review, 52 (2010), pp. 471–501.
[45] J. Renegar and B. Grimmer, *A simple nearly optimal restart scheme for speeding up first-order methods*, Foundations of Computational Mathematics, 22 (2022), pp. 211–256.
[46] R. T. Rockafellar, *Convex analysis*, vol. 11, Princeton university press, 1997.
[47] V. Roulet, N. Boumal, and A. d'Aspremont, *Renegar's condition number, sharpness and compressed sensing performance*, arXiv preprint arXiv:1506.03295, (2015).
[48] V. Roulet and A. d'Aspremont, *Sharpness, restart, and acceleration*, SIAM Journal on Optimization, 30 (2020), pp. 262–289.
[49] M. Rudelson and R. Vershynin, *Sparse reconstruction by convex relaxation: Fourier and gaussian measurements*, in 2006 40th Annual Conference on Information Sciences and Systems, IEEE, 2006, pp. 207–212.
[50] F. Shang, J. Cheng, Y. Liu, Z.-Q. Luo, and Z. Lin, *Bilinear factor matrix norm minimization for robust pca: Algorithms and applications*, IEEE transactions on pattern analysis and machine intelligence, 40 (2017), pp. 2066–2080.
[51] J. F. Sturm, *Error bounds for linear matrix inequalities*, SIAM Journal on Optimization, 10 (2000), pp. 1228–1248.
[52] R. Vershynin, *Introduction to the non-asymptotic analysis of random matrices*, Cambridge University Press, 2012, p. 210–268.
[53] T. Yang and Q. Lin, *Rsg: Beating subgradient method without smoothness and strong convexity*, The Journal of Machine Learning Research, 19 (2018), pp. 236–268.
[54] S. Zhang, *Global error bounds for convex conic problems*, SIAM Journal on Optimization, 10 (2000), pp. 836–851.
[55] Z. Zhou and A. M.-C. So, *A unified approach to error bounds for structured convex optimization problems*, Mathematical Programming, 165 (2017), pp. 689–728.