Contents

1 Notes on Linear Inverse Problems without RIP

This is a collections of notes centered around "Mathematics of sparsity (and a few other things)" from Emmanuel Candes [4]. I try to go a bit deeper into the derivation of the results and the intuition behind them than he does. The original paper only covers RIPless theory in the noiseless case, so we'll mostly restrict ourselves to that context. We will also focus on the compressed sensing problem since that is the most broadly discussed form of linear inverse problem.

Work on linear inverse problems explores conditions under which it is possible to recover a signal from linear measurements. Of particular interest is the case where the number of measurements is far fewer than the size of the ambient dimension of the target signal. The name "compressed sensing" arises from essentially that: rather than measure a signal and then compressing it to just the required information, instead, directly measure the compressed signal. More generally, we would like to recover a high-dimensional signal that has some low-dimensional structure (and thus is in some sense compressible). As discussed towards the end of [4], this theory goes pretty far and applies to problems such as source separation and demixing, robust PCA, phase retrieval, and matrix completion. Examples of low-dimensional structure include sparsity, low-rankness. Chandrasekaran et al. give a more exhaustive list in section 2.2 of [3].

1.1 Compressed sensing

Consider the ℓ_1 minimization problem in $\mathbb{R}^d$:

$$\min_x \|x\|_1 \text{ subject to: } Ax = y$$

We wish to recover an unknown vector $\bar{x}$ given a set of linear measurements of the form $y_i = \langle a_i, \bar{x} \rangle$. It's probably surprising that if $\bar{x}$ is an s -sparse vector, and the measurement vectors a_i are sampled from an isotropic distribution (identity covariance matrix), we only need $O(s \log d)$ measurements to exactly recover $\bar{x}$ from y by solving the ℓ_1 minimization problem above. It's immediately clear that low-dimensional structure is not sufficient to ensure recovery of the unknown signal: we must also choose our sampling vectors a_i well.

1.1.1 Conditions on the sampling vectors

Note that the isotropy condition on the distribution of the sampling matrix ensures that the matrix $1/n \sum a_i a_i^*$ (where $\# \cdot$ ^*$ denotes the conjugate-transpose) converges to the identity, meaning that the measurement matrix has a left inverse given enough measurements. Thus, any x is recoverable given sufficiently many samples. In constrast, without this condition, there would exist signals x that are not recoverable even with infinitely many samples (they lie in the nullspace of A in expectation). See the discussion in section 1.3 of [2].

But isotropy of the sensing matrix with low-dimensional structure of the signal of interest is also not enough for recovery. Consider recovering an s-sparse signal using sampling vectors that are uniformly chosen from unit vectors along the coordinate axes. In this case, each sample reveals information about only a single entry, so we would require (as in the coupon collector problem) $d \log d$ samples to recover all the entries. In order to provide the rate of $O(s \log d)$ presented earlier, we need a bound on the "coherence" of the sampling vectors. Continuing with case of compressed sensing, we call $\mu(F)$ the smallest value larger than the squared magnitude of the entries of $a \sim F$ (either whp or wp 1). For sensing distributions F that have the isotropy property (identity covariance) this gives a smallest possible coherence μ(F) of 1. In the case described earlier that displays the coupon collector behavior, we have a coherence of d, which is of course large since our rate is no longer logarithmic in the ambient dimension. Lighter tailed distributions lead to a lower coherence than heavier tailed distributions (see the discussion after equation 1.7 and in section 1.3 from [2]).

1.1.2 Certifying Optimality and Uniqueness

To understand where these conditions come from, let's look at the KKT conditions for the ℓ_1 minimization problem. A feasible point is optimal if we can find a dual certificate.

First, forming the Lagrangian:

$$\mathcal{L}(x, \lambda) = \|x\|_1 + \lambda^*(Ax - y)$$

Applying first order optimality:

$$\partial \|x\|_1 + A^* \lambda = 0$$

So if there exists some vector $u \in range(A^*)$ that is also a subgradient of the ℓ_1 norm at x, then x is optimal. Notice that range(A*) $\perp$ null(A). Also, recall that $\partial \|x\|_1 = w | w_i = sign(x)$ if $x_i > 0, w_i \in [-1, 1] \, otherwise.$ So we obtain a dual certificate by finding some vector v $\perp$ null(A) that is sign(x$_\mathrm{i}$) on the support of x, and smaller than 1 in magnitude off the support of x.

One way to arrive at such a v is to consider the "ansatz" problem introduced in section 4 of [4]:

min $\|v\|_2$ subject to: $v \perp null(A), v_i = sign(x_i)$ for i in the support of x, $v_i \in (-1, 1)$ for i off the support of x.

The solution for which can be found in closed form given that the nullspace of A has only a trivial intersection with the support of x. In fact, the solution to the above problem certifies more than optimality (notice (-1, 1) in the constraint compared to [-1, 1] from the subdifferential above), it also certifies that x is the unique solution:

Following theorem 6.8 from Yuxin Chen's lecture notes for ELE 538B:

Let T be the set of nonzero indices of x, let $\mathrm{A_T}$ be A with 0 in columns not in T (the restriction of A to the support of x). Assume further that $A_T^* A_T$ is invertible. Consider x, an optimal solution to the ℓ_1 minimization problem and a displacement vector h in the nullspace of A. By convexity of the feasible set, if z is another optimal solution, it must be of the form x + h.

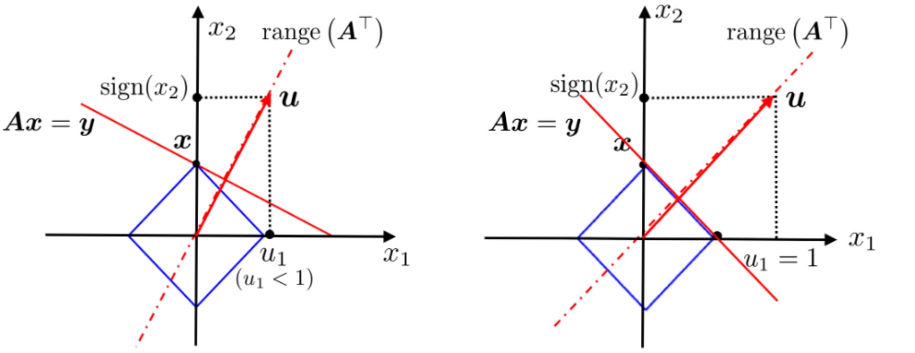
Let w be a subgradient of the ℓ_1 norm at x:

$$\|x + h\|_1 \geq \|x\|_1 + \langle g, h \rangle$$

Rewriting $\langle$ w, h $\rangle$ = $\langle$ v, h $\rangle$ + $\langle$ w - v, h $\rangle$ = $\langle$ A^*u, h $\rangle$ + $\langle$ w - v, h $\rangle$ for some u since v is in the range of A^*. The first term is 0 since h is in the nullspace of A. Now, we can choose w such that $\mathrm{w_i} = \mathrm{sign}(\mathrm{x_i})$ if $\mathrm{i} \in \mathrm{T}$, and $\mathrm{sign}(\mathrm{h_i})$ otherwise. In that case, the second term reduces to:

$$\langle w - v, h \rangle = \sum_{i \notin T} (sign(h_i) - v_i) h_i = \sum_{i \notin T} |h_i| - v_i h_i$$

But this is strictly greater than 0 unless $\mathrm{h_i}$ is 0 off the support of x. In that case, Ah = $\mathrm{A_T h_T}$ = 0 since h is in the nullspace of A. But $\mathrm{A_T}$ has full column rank, so $\mathrm{h_T}$ = 0 otherwise we have a contradiction. Putting it all together, we have $\|x + h\|_1 > \|x\|_1$ for all h in the nullspace of a, so x is the unique optimum.



When $|u_1| < 1$, solution is unique   When $|u_1| = 1$, solution is non-unique

Section 4 of [4] shows how the equality constraint $\mathrm{v_i} = \mathrm{sign}(\mathrm{x_i})$ can be loosened to hold approximately. The so-called "golfing" scheme then gives an iterative process for computing this approximate solution which can be shown to exist with high probability given the isotropy condition on the rows of A. See the proof of lemma 3.3 from [2].

The matrix completion literature has analogous results, as discussed in [4].

1.2 Gaussian Models and Phase Transitions

Stated more generally than in the previous section, x is the unique solution if and only iff the nullspace of A has only a trivial intersection with the directions that decrease the ℓ_1 norm at x. In other words x, is the unique optimum iff $null(A) \cap \mathcal{T}(\|\cdot\|_1, x) = \{0\}$ where $\mathcal{T}(\|\cdot\|_1, x)$ is the tangent cone (or cone of descent) of the ℓ_1 norm at x. Although straightforward to show, see the proof of proposition 2.1 in [3].

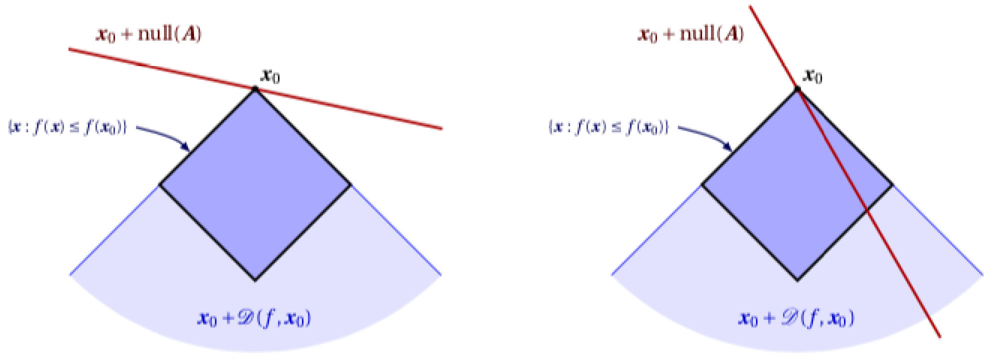
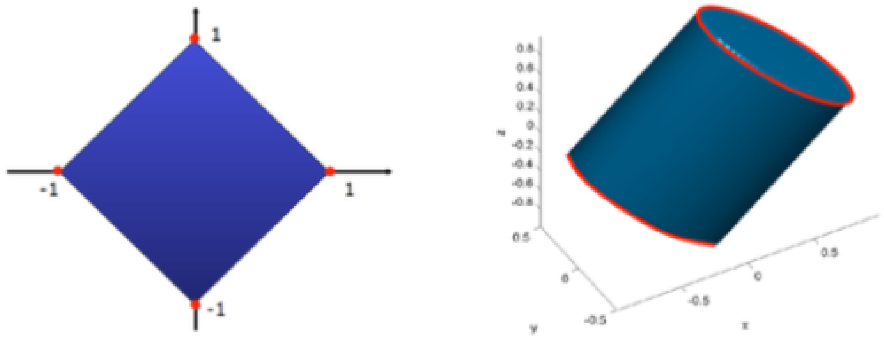

$x_0 + \text{null}(A)$
x_0
$\{x : f(x) \leq f(x_0)\}$
$x_0 + \mathcal{D}(f, x_0)$
$x_0 + \text{null}(A)$
x_0
$\{x : f(x) \leq f(x_0)\}$
$x_0 + \mathcal{D}(f, x_0)$

FIGURE 2.3: **The optimality condition for a regularized inverse problem.** The condition for the regularized linear inverse problem (2.4) to succeed requires that the descent cone $\mathcal{D}(f, x_0)$ and the null space null(A) do not share a ray. **[left]** The regularized linear inverse problem succeeds. **[right]** The regularized linear inverse problem fails.

The ℓ_1 (ball shown in a) and nuclear (ball for symmetric 2x2 matrices shown in b) norm have small tangent cones at sparse points, which explains why we arrive at sparse solutions:



(a)

(b)

Gordon's "escape through the mesh" lemma bounds the probability that a uniformly distributed subspace misses a convex cone [5]. We can use this result to provide sample requirements for recovery success whp for the gaussian case. Stating a corollary (3.3 in [3]) of the theorem:

For a convex cone C $\in \mathbb{R}^d$ and an n x d gaussian map A, null(A) $\cap$ C $=$ $\{0\}$ with high probabily if n (number of samples) $>=$ w(C)^2 $+$ 1 where w(C) is the gaussian width of the cone C.

The gaussian width of a set C is given by: $w(C) = \mathbb{E}sup_{z \in C \cap S^{d-1}} g^* z$ where g $\sim$ N(0, 1) and S^{d-1} is the unit sphere in d dimensions. Notice that this is a expected supremum of a gaussian process and can be bound by Dudley's inequality, but computing covering numbers for convex cones is difficult (see theorem 3.5 of [3]).

With this result, we can give a bound on the required number of samples for recovery (see proposition 3.10 in [3]):

Let $\bar{x}$ be an s-sparse vector in $\mathbb{R}^d$. Recovery by ℓ_1 minimization succeeds whp if we have at least $2s\ \log(d/s) + 5/4s + 1$ random gaussian samples.

The proof proceeds by bounding the gaussian width:

Apply weak duality to show that $w(C) \leq \mathbb{E}dist(g, C^{\circ})$ where dist is the euclidean distance and Cˆˆ is the polar cone of C. Since we seek to bound the gaussian width of the ℓ_1 tangent cone at , this shows that we only need to bound the expected distance of a gaussian vector to the ℓ_1 normal cone at . The normal cone is simply the conic hull of the subdifferential at x.

Continuing with the application of the corollary, we want to bound w(C)^2, so applying Jensen's:

$$w(C) \leq Edist(g, \partial\|\bar{x}\|_1)^2 = \mathbb{E}\inf_{w} \sum (g_i - w_i)^2 forw \in cone\partial\|\bar{x}\|_1$$

Breaking up the sum:

$$\mathbb{E}\inf_{w} \sum (g_i - w_i)^2 = E \inf_{t, |w_i| \leq 1} \sum_{i \in T} (g_i - sign(\bar{x}_i)t)^2 + \sum_{i \notin T} (g_i - sign(w_i)t)^2$$

Where we introduce t since we consider w in the conic hull of the subdifferential.

Then for any t, the first term is at most s(1 $+$ t^2) (by taking expectation). The second term is bound by applying integration by parts, applying the bound on the gaussian hazard function: 1/x ϕ(x) and then finally optimizing to minimize the upper bound and noting that $(1 - s/d)/(\pi\sqrt{log(d/s)} < 1/4$. See Appendix C of [3] for details.

1.2.1 Phase Transitions in the Gaussian case

Earlier, we saw that the we require an incoherent and isotropic sampling distribution. In the special case of a gaussian sampling matrix, the nullspace of A is uniformly distributed over the subspaces of dimension $d - m$ in $\mathbb{R}^d$ (meaning that the nullspace is distributed as a random rotation of a d-m subspace).

Consider a subspace M of size m and another randomly oriented subspace N of size n in d dimensions. The probability that the subspaces have a non-trivial intersection is 1 if m $+$ n $>$ d and 0 otherwise. Analogously to the sharp phase transition seen here, we also see sharp phase transitions for the intersection of randomly oriented convex cones.

The recovery bounds discussed previously arise from bounds on the probability that a cone (the cone of descent of the objective) has a non-trivial intersection with a subspace (the nullspace of our sensing matrix). Thus, we see a similar phase transition for the probability of successful recovery. See [1].

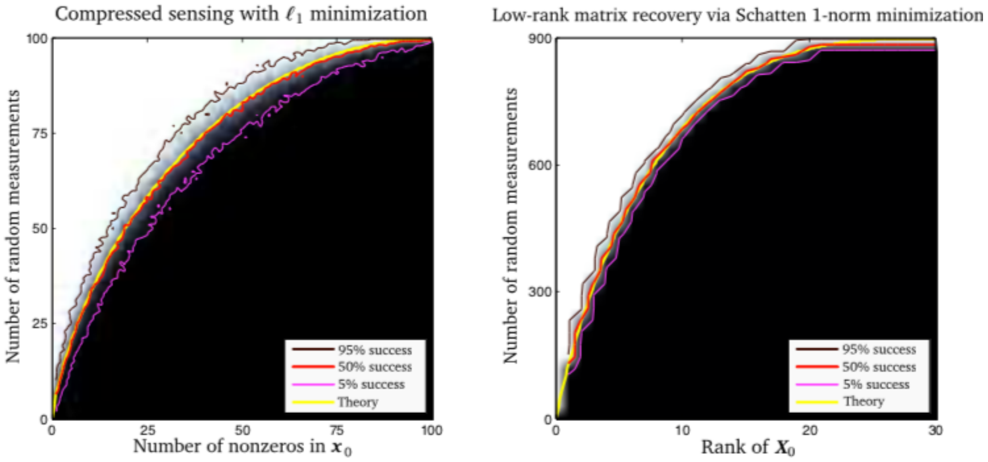


FIGURE 2.4: **Phase transitions for regularized linear inverse problems. [left] Recovery of sparse vectors.** The empirical probability that the ℓ_1 minimization problem (2.5) identifies a sparse vector $x_0 \in \mathbb{R}^{100}$ given random linear measurements $z_0 = Ax_0$. **[right] Recovery of low-rank matrices.** The empirical probability that the S_1 minimization problem (2.6) identifies a low-rank matrix $X_0 \in \mathbb{R}^{30\times 30}$ given random linear measurements $z_0 = \mathscr{A}(X_0)$. In each panel, the heat map indicates the empirical probability of success (black = 0%; white = 100%). The yellow curve marks the theoretical prediction of the phase transition from Theorem II; the red curve traces the 50% success isocline calculated from the data.

1.3 No RIP

Most of the work on compressed sensing relies on some condition similar to RIP. The work discussed here focuses on results obtained with conditions that are easier to verify than RIP (refer to the definition of RIP and it will be clear that it is difficult to verify). In constrast, RIP is a uniform condition (holds for all x) while the results discussed here apply only to a fixed x. Essentially, unlike in the RIP case, a given sampling matrix A can recover a fixed x with high probability, but that same A cannot be used to recovery arbitrary x. See the discussion in section 1.7 of [2].

1.4 Questions

- The rates given in theorem 1 of [4] are tight up to a constant factor in the sense that there exist signals such that given fewer than $\mu\, s\, logn$ samples, recovery is impossible. How to construct such a signal?

- Do we see similar phase transitions with non-gaussian sensing matrices?

- Is the nullspace of other non-gaussian maps also uniformly distributed over subspaces in the codimension?

1.5 References

[1] D. Amelunxen, M. Lotz, M. B. McCoy, and J. A. Tropp, "Living on the edge: Phase transitions in convex programs with random data," Arxiv, 2013.

[2] E. J. Candes and Y. Plan, "A probabilistic and RIPless theory of compressed sensing," Arxiv, 2010.

[3] V. Chandrasekaran, B. Recht, P. A. Parrilo, and A. S. Willsky, "The Convex Geometry of Linear Inverse Problems," Arxiv, 2010, doi: 10.1007/s10208-012-9135-7.

[4] Candès, Emmanuel J. "Mathematics of sparsity (and a few other things)." Proceedings of the International Congress of Mathematicians, Seoul, South Korea. Vol. 123. Citesee, 2014.

[5] Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in Rn. Springer, 1988.