1 The problem

Linear regression \( y^i = \theta^T x^i + \epsilon^i \) for \( i = 1 : n \) with \( \theta, x^i \in \mathbb{R}^p \), and \( \epsilon^i \) some zero-mean i.i.d. noise. In matrix form, the problem is

\[
    y = X\theta + \epsilon \quad X \in \mathbb{R}^{n \times p}, \ y \in \mathbb{R}^n, \ \theta \in \mathbb{R}^p. \tag{1}
\]

Now, let \( p \ll n \), that is, there are more parameters to estimate than samples. In this case, (1) has infinitely many solutions.

Estimating \( \theta \) is possible, and efficient, with the assumption that \( \theta \) is a sparse vector, i.e. that \( ||\theta||_0 \leq s \ll p \). The notation \( ||v||_0 \) denotes the number of non-zero elements in the vector \( v \). One also denotes \( ||v||_p = \left( \sum_i |v_i|^p \right)^{\frac{1}{p}} \) for \( 0 < p < 1 \), even if these are not norms (they don’t satisfy triangle inequality)\(^1\). Vectors with small 0-norm are called sparse, vectors with small \( p \)-norm (\( p < 1 \)) are called compressible.

\(^1\)Note that \( \lim_{p \to 0} ||x||_p = \infty \) when \( x \neq 0 \), so the zero norm is not the limit of the \( p \)-norm.
2 Solving (1)

Suprisingly, if $\theta$ is sparse, it can be estimated by a convex program. For example, the Danzig selector is the program

$$\min_{\theta} ||\theta||_1 \text{ s.t. } ||X^T(X\theta - y)||_2 \leq \delta$$

where $\delta$ is a parameter proportional to the noise variance or amplitude. This is a convex optimization problem with $p$ unknowns and 1 constraint. A variant for bounded noise is

$$\min_{\theta} ||\theta||_1 \text{ s.t. } ||X\theta - y||_\infty \leq \delta,$$

Yet another variant recalls the well-known Lasso estimator for linear regression.

$$\min_{\theta} ||\theta||_1 \text{ s.t. } ||X\theta - y||_2 \leq \delta,$$

The problem (3) can be turned into a linear program with variable $t^+_i = \max\{\theta, 0\}$, $t^-_i = \max\{-\theta, 0\}$

$$\min_t 1^T t \text{ s.t. } t \geq 0, \|[X - X] \begin{bmatrix} t^+ \\ t^- \end{bmatrix} - y\| \leq \delta$$

What can we say about the solutions to these linear programs?

**Theorem 1** If there is an $x^*$ such that $y = Ax^* + e$ with $||e||_2 \leq \epsilon$, and if $n \geq C_s\log(p/s)$, $\delta_{2s}(A) \leq \sqrt{2} - 1$, then

$$||\hat{x} - x^*||_2 \leq \frac{c_0}{\sqrt{s}} ||x^* - x^*_s||_1 + c_1\epsilon$$

where $\hat{x}$ is the solution to optimization problem (4) and $x^*_s$ is the best $s$-sparse approximation to $x^*$.

A few observations

1. The best $s$-sparse approximation to a vector $x$ is obtained as follows
(a) sort $|x_i|$ (the elements of $x$) in decreasing order. Denote the $j$ largest magnitude element by $|x[^j]|$.

(b) 

$$(x_s)_i = \begin{cases} 
  x_i & \text{if } i \in \{[1], \ldots, [s]\} \\
  0 & \text{otherwise}
\end{cases}$$

In words, $x_s$ zeroes out all but the $s$ largest in magnitude elements of $x$.

2. Hence, if $x^*$ is $s$-sparse, then $||x^* - x_s||_1 = 0$, if $x^*$ is compressible, the residual $||x^* - x_s||$ is small.

3. The condition $||e||_2^2 \leq \epsilon$ can be replaced with the statistical condition $Var e \leq \frac{c \epsilon}{n}$ with $c < 1$ a constant

4. The first term in (6) is proportional to the “noise” in the inputs, i.e. the difference between $x^*$ and the closest sparse vector to it; the second term is proportional to the measurement noise. When both terms are small, we are guaranteed that the convex approximation gives a good solution to our original, non-convex, problem.

**Restricted Isometry of $A$.** The one important condition of the theorem is on the measurement matrix $A$. We define $\delta_s(A)$ to be the smallest positive number $\delta$ satisfying

$$
(1 - \delta)||x||^2 \leq ||Ax||^2 \leq (1 + \delta)||x||^2 \text{ for all } s\text{-sparse vectors } x
$$

Condition (7) is called the **Restricted Isometry Property (RIP)**. Essentially, RIP implies that even though $A$ as a whole cannot be orthogonal matrix, any subset of $s$ or fewer columns from $Z$ forms an approximately orhtogonal matrix.

**Exercise:** What is the largest value of $1 + \delta$? What is the smallest? (Assume that the columns of $A$ have norm 1.)

Theorem 1 requires that $\delta_{2s}(A)$ be small. This condition essentially ensures that the linear system $Ax = y$ has a unique (approximate) sparse solution. Assume for the moment that there were two $s$-sparse vectors $x, x'$ satisfying $Ax = y$. Then, $A(x - x') = 0$ and $x - x'$ is a $2s$-sparse vector in the null space of $A$. By the RIP condition of Theorem 1, there cannot be any $2s$ sparse vector $z$ so that $Az = 0$ except for $z = 0$.

Hence, intuitively, the $l_1$ minimization assures that we find a sparse solution,
and the RIP property ensures that the solution found is unique, and therefore must be (approximately) equal to the true $x^*$.

**Minimum number of samples and information limit of sparse recovery.** The information needed to find the $s$ non-zero elements of $x$ is the entropy of a uniform prior over all $s$-subsets of $\{1:p\}$. Hence $\#\text{bits} = \log\binom{p}{s}$.

To approximate this number we use Striling’s approximation $n! \approx \sqrt{2\pi n} \frac{n^n}{e^n}$.

We obtain

$$\#\text{bits} \approx \log \frac{p^p \cdot e^s \cdot (p-s)^{p-s}}{e^p s^s (p-s)^{p-s}} = p \log p - s \log s - (p-s) \log (p-s)$$

$$> p \log p - s \log s - (p-s) \log p = s \log \left(\frac{p}{s}\right)$$

(9)

Thus the number of independent measurement needed to obtain a sparse $x^*$ by **by any method whatsoever** is only a constant\(^2\) smaller than the number of samples required by Compressed Sensing (aka Theorem 1).

### 3 What matrices $A$ satisfy the RIP?

RIP as defined here is only one version of a spectrum of similar properties; sometimes these properties are known alternatively as *incoherence* properties of $A$.

Verifying the RIP for a given matrix requires examining $\binom{p}{2s}$ submatrices, therefore it is intractable in practice. This is unfortunately true no matter which of the many incoherence properties one uses.

Therefore, in the vast majority of research on compressed sensing, people resort to constructing $A$ matrices which can be *proved* to have the RIP.

\(^2\)This constant is not large, e.g 4.
One easy way to obtain such an $A$ is by sampling. For example

\[
\begin{align*}
A_i & \sim \text{random unit vector in } \mathbb{R}^p \\
A_i & \sim \text{random vector of the Fourier basis } \in \mathbb{R}^p \\
A_{ij} & \sim \text{Normal}(0, \frac{1}{n}) \\
A_{ij} & \sim \frac{1}{\sqrt{n}} \text{uniform}(\pm 1)
\end{align*}
\]

all produce, with probability of success at least $1 - e^{-cn}$ ($c$ depending on the particular sampling method), matrices satisfying RIP. The coefficients in the last two generative processes ensure that the columns of $A$ are unit vectors.

Deterministic constructions of $A$ are more complicated, and the existing ones use either results from the theory of codes, or from the theory of expander graphs.\(^3\)

### 4 Iterative algorithms for compressed sensing

TBW. See STAT593C Lecture 9.

### 5 Summary

In summary, we can find the sparse solution of $Ax = y$ with $p$ unknowns and $n \ll p$ equations if:

- the solution is $s$ sparse,
- we have order $s \log(p/s)$ measurements
- the measurement matrix has no sparse vectors in its null-space

---

\(^3\)An expander graph is a graph for which any “small” subset has a “large” set of neighbors. In other words, a Markov Chain would mix well on an expander graph. As it turns out, most graphs are expander graphs; yet, for a given graph, it is hard to verify that the graph is an expander. See e.g. [http://www.math.ias.edu/~boaz/ExpanderCourse/](http://www.math.ias.edu/~boaz/ExpanderCourse/)
With these conditions satisfied

- the system can be solved efficiently (by LP, convex optimization, iterative projection methods, etc)
- the system can be solved in the presents of noise in the measurements ($e$) and in the inputs ($x$ compressible but not sparse)

This is an area of intense current activity, and the principles of sparse recovery have been expanded to recovery of sparse matrices, low rank matrices, and so on.