Convex Optimization M2

$\ell_1$-recovery, compressed sensing
Today

- Sparsity, low complexity models.
- $\ell_1$-recovery results: three approaches.
- Extensions: matrix completion, atomic norms.
- Algorithmic implications.
Consider the following underdetermined linear system

\[
A \ x = b
\]

where \( A \in \mathbb{R}^{m \times n} \), with \( n \gg m \).

Can we find the \textit{sparsest} solution?
Introduction

- **Signal processing:** We make a few measurements of a high dimensional signal, which admits a sparse representation in a well chosen basis (e.g. Fourier, wavelet). Can we reconstruct the signal exactly?

- **Coding:** Suppose we transmit a message which is corrupted by a few errors. How many errors does it take to start losing the signal?

- **Statistics:** Variable selection in regression (LASSO, etc).
Introduction

Why sparsity?

- Sparsity is a proxy for power laws. Most results stated here on sparse vectors apply to vectors with a power law decay in coefficient magnitude.
- Power laws appear everywhere . . .
  - Zipf law: word frequencies in natural language follow a power law.
  - Ranking: pagerank coefficients follow a power law.
  - Signal processing: $1/f$ signals
  - Social networks: node degrees follow a power law.
  - Earthquakes: Gutenberg-Richter power laws
  - River systems, cities, net worth, etc.
Introduction

Frequency vs. word in Wikipedia (from Wikipedia).

A. d'Aspremont. M1 ENS.
Introduction

Frequency vs. magnitude for earthquakes worldwide. Christensen et al. [2002]

Frequency vs. magnitude for earthquakes worldwide. Christensen et al. [2002]
Pages vs. Pagerank on web sample. Pandurangan et al. [2006]
Getting the sparsest solution means solving

\[
\begin{align*}
\text{minimize} & \quad \text{Card}(x) \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

which is a (hard) \textbf{combinatorial} problem in \( x \in \mathbb{R}^n \).

A classic heuristic is to solve instead

\[
\begin{align*}
\text{minimize} & \quad \|x\|_1 \\
\text{subject to} & \quad Ax = b
\end{align*}
\]

which is equivalent to an (easy) \textbf{linear program}.
Introduction

Assuming $|x| \leq 1$, we can replace:

$$
\text{Card}(x) = \sum_{i=1}^{n} 1_{\{x_i \neq 0\}}
$$

with

$$
\|x\|_1 = \sum_{i=1}^{n} |x_i|
$$

Graphically, assuming $x \in [-1, 1]$ this is:

The $l_1$ norm is the **largest convex lower bound** on $\text{Card}(x)$ in $[-1, 1]$. 

A. d’Aspremont. M1 ENS.
Example: we fix $A$, we draw many sparse signals $e$ and plot the probability of perfectly recovering $e$ by solving

$$
\begin{align*}
\text{minimize} & \quad \|x\|_1 \\
\text{subject to} & \quad Ax = Ae
\end{align*}
$$

in $x \in \mathbb{R}^n$, with $n = 50$ and $m = 30$. 

![Plot showing probability of recovering $e$ vs. cardinality of $e$.]
Donoho and Tanner [2005] and Candès and Tao [2005] show that for certain classes of matrices, when the solution $e$ is sparse enough, the solution of the $\ell_1$-minimization problem is also the sparsest solution to $Ax = Ae$.

Let $k = \text{Card}(e)$, this happens even when $k = O(m)$ asymptotically, which is provably optimal.

Also obtain bounds on reconstruction error outside of this range.
$l_1$ recovery
Kashin and Temlyakov [2007]: Simple relationship between the diameter of a section of the $\ell_1$ ball and the size of signals recovered by $\ell_1$-minimization.

**Proposition**

**Diameter & Recovery threshold.** Given a coding matrix $A \in \mathbb{R}^{m \times n}$, suppose that there is some $k > 0$ such that

$$\sup_{A x = 0, \|x\|_1 \leq 1} \|x\|_2 \leq \frac{1}{\sqrt{k}} \quad (1)$$

then sparse recovery $x^{LP} = u$ is guaranteed if Card$(u) \leq k/4$, and

$$\|u - x^{LP}\|_1 \leq 4 \min_{\{\text{Card}(y) \leq k/16\}} \|u - y\|_1$$

where $x^{LP}$ solves the $\ell_1$-minimization LP and $u$ is the true signal.
Proof. Kashin and Temlyakov [2007]. Suppose
\[
\sup_{\|x\|_1 \leq 1} \|Ax\|_2 \leq k^{-1/2}
\]

Let \( u \) be the true signal, with \( \text{Card}(u) \leq k/4 \). If \( x \) satisfies \( Ax = 0 \), for any support set \( \Lambda \) with \( |\Lambda| \leq k/4 \),
\[
\sum_{i \in \Lambda} x_i \leq \sqrt{|\Lambda|} \|x\|_2 \leq \sqrt{|\Lambda|/k} \|x\|_1 \leq \|x\|_1/2,
\]
Now let \( \Lambda = \text{supp}(u) \) and let \( v \neq u \) such that \( x = v - u \) satisfies \( Ax = 0 \), then
\[
\|v\|_1 = \sum_{i \in \Lambda} |u_i + x_i| + \sum_{i \notin \Lambda} |x_i| \geq \sum_{i \in \Lambda} |u_i| - \sum_{i \in \Lambda} |x_i| + \sum_{i \notin \Lambda} |x_i| = \|u\|_1 + \|x\|_1 - 2 \sum_{i \in \Lambda} |x_i|
\]
and
\[
\|x\|_1 - 2 \sum_{i \in \Lambda} |x_i| > 0
\]
means that \( \|v\|_1 > \|u\|_1 \), so \( x^{LP} = u \). The error bound follows from similar arg.
Diameter, low $M^*$ estimate

**Theorem**

**Low $M^*$ estimate.** Let $E \subset \mathbb{R}^n$ be a subspace of codimension $k$ chosen uniformly at random w.r.t. to the Haar measure on $G_{n,n-k}$, then

$$\text{diam}(K \cap E) \leq c \sqrt{\frac{n}{k}} M(K^*) = c \sqrt{\frac{n}{k}} \int_{S^n_{n-1}} \|x\|_{K^*} d\sigma(x)$$

with probability $1 - e^{-k}$, where $c$ is an absolute constant.

**Proof.** See [Pajor and Tomczak-Jaegermann, 1986] for example.

We have $M(B^n_{\infty}) \sim \sqrt{\log n/n}$ asymptotically. This means that random sections of the $\ell_1$ ball with dimension $n - k$ have diameter bounded by

$$\text{diam}(B^n_1 \cap E') \leq c \sqrt{\frac{\log n}{k}}$$

with high probability, where $c$ is an absolute constant (a more precise analysis allows the log term to be replaced by $\log(n/k)$).
Sections of the $\ell_1$ ball

The bound $\text{diam}(B_1^n \cap E) \leq c \sqrt{\frac{\log n}{k}}$ means recovery of all signals with at most

$$O\left(\frac{k}{\log n}\right)$$

coefficients, using $k$ linear observations $Ae$.

Results guaranteeing near-optimal bounds on the diameter can be traced back to Kashin and Dvoretzky’s theorem.

- **Kashin decomposition** [Kashin, 1977]. Given $n = 2m$, there exists two orthogonal $m$-dimensional subspaces $E_1, E_2 \subset \mathbb{R}^n$ such that

$$\frac{1}{8}\|x\|_2 \leq \frac{1}{\sqrt{n}}\|x\|_1 \leq \|x\|_2, \quad \text{for all} \ x \in E_1 \cup E_2$$

- In fact, most $m$-dimensional subspaces satisfy this relationship.
Similar results exist for rank minimization.

- The $\ell_1$ norm is replaced by the trace norm on matrices.
- Exact recovery results are detailed in Recht et al. [2007], Candes and Recht [2008], . . .
References


