Optimisation Combinatoire et Convexe.

Low complexity models, $\ell_1$ penalties.
Today

- Sparsity, low complexity models.
- \(1\)-recovery results: three approaches.
- Extensions: matrix completion, atomic norms.
- Algorithmic implications.
Consider the following underdetermined linear system

\[ A \times x = b \]

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

Can we find the **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).
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).
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) 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) linear program.
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 $[-1, 1]$. 
Example: we fix $A$, we draw many sparse signals $e$ and plot the probability of perfectly recovering $e$ by solving

\[
\begin{align*}
\text{minimize} & \quad kx_k \\
\text{subject to} & \quad Ax = Ae
\end{align*}
\]

in $x \in \mathbb{R}^n$, with $n = 50$ and $m = 30$. 
Donoho and Tanner [2005] and Candès and Tao [2005] show that for certain classes of matrices, when the solution \( \mathbf{e} \) is sparse enough, the solution of the \( \ell_1 \)-minimization problem is also the \textbf{sparsest} solution to \( \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{e} \).

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

Also obtain bounds on reconstruction error outside of this range.
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], ...
\[ 1 \] \text{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.** \( \text{Given a coding matrix } A \in \mathbb{R}^{m \times n}, \text{ suppose that there is some } k > 0 \text{ such that} \)

\[
\sup_{A\mathbf{x}=0, \|\mathbf{x}\|_1 \leq 1} k\|\mathbf{x}\|_2 \leq \frac{1}{k} \tag{1}
\]

*then sparse recovery \( \mathbf{x}_{\text{LP}} = \mathbf{u} \) is guaranteed if \( \text{Card}(\mathbf{u}) \leq k/4 \), and*

\[
ku - \mathbf{x}_{\text{LP}} \leq 4 \min_{\{\text{Card}(\mathbf{y}) \leq k/16\}} ku - y\mathbf{k}_1
\]

*where \( \mathbf{x}_{\text{LP}} \) solves the \( \ell_1 \)-minimization LP and \( \mathbf{u} \) is the true signal.*
**Diameter**

**Proof.** Kashin and Temlyakov [2007]. Suppose

\[
\sup_{\|x\|_1 \leq 1} k \|x\|_2 \leq k^{-1/2}
\]

Let \( u \) be the true signal, with \( \text{Card}(u) \leq k/4 \). If \( x \) satisfies \( Ax = 0 \), for any support set \( \Lambda \) with \( |\Lambda| \leq k/4 \),

\[
\sum_{i \in \Lambda} x_i \leq \sqrt{|\Lambda|} k \|x\|_2 \leq \sqrt{|\Lambda|/k} k \|x\|_1 \leq k \|x\|_1/2,
\]

Now let \( \Lambda = \text{supp}(u) \) and let \( v \not= u \) such that \( x = v - u \) satisfies \( Ax = 0 \), then

\[
k \|v\|_1 = \sum_{i \in \Lambda} |u_i + x_i| + \sum_{i \not\in \Lambda} |x_i| \geq \sum_{i \in \Lambda} |u_i| - \sum_{i \in \Lambda} |x_i| + \sum_{i \not\in \Lambda} |x_i| = k \|u\|_1 + k \|x\|_1 - 2 \sum_{i \in \Lambda} |x_i|
\]

and

\[
k \|x\|_1 - 2 \sum_{i \in \Lambda} |x_i| > 0
\]

means that \( k \|v\|_1 > k \|u\|_1 \), so \( x^{LP} = u \). The error bound follows from similar arg.
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

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

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

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

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

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

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

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

$$\frac{1}{8} \|x\|_2 \leq \frac{1}{\sqrt{n}} \|x\|_1 \leq \|x\|_2,$$

for all $x \in E_1 \cup E_2$.

- In fact, most $m$-dimensional subspaces satisfy this relationship.
Atomic norms

We can give another geometric view on the recovery of low complexity models. Once again, we focus on problem (2), namely

\[
\begin{align*}
\text{minimize} & \quad \|x\|_A \\
\text{subject to} & \quad Ax = Ax_0
\end{align*}
\]

and we start by a construction from [Chandrasekaran et al., 2010] on a specific type of norm penalties, which induce simple representations in a generic setting.

**Definition**

**Atomic norm.** Let \( A \subset \mathbb{R}^n \) be a set of atoms. Let \( k \cdot k_A \) be the gauge of \( A \), i.e.

\[
k x k_A = \inf \{ t > 0 : x \in t \times \text{Co}(A) \}
\]

The motivation for this definition is simple, if the centroid of \( A \) is at the origin, we have

\[
k x k_A = \inf \left\{ \sum_{a \in A} \lambda_a : x = \sum_{a \in A} \lambda_a a, \lambda_a \geq 0 \right\}
\]
Atomic norms

Depending on $A$, atomic norms look very familiar.

- **Suppose** $A = \{\pm e_i\}_{i=1,...,n}$ where $e_i$ is the Euclidean basis of $\mathbb{R}^n$. Then $\text{Co}(A)$ is the $\ell_1$ ball and $kxk_A = kxk_1$.

- **Suppose** $A = \{uv^T : u, v \in \mathbb{R}^n, ku_k_2 = kv_k_2 = 1\}$, then $\text{Co}(A)$ is the unit ball of the trace norm and $kXk_A = kXk_*$ when $X \in \mathbb{R}^{n \times n}$.

- **Suppose** $A$ is the set of all orthogonal matrices of dimension $n$. Its convex hull is the unit ball of the spectral norm, and $kXk_A = kXk_2$ when $X \in \mathbb{R}^{n \times n}$.

- **Suppose** $A$ is the set of all permutations of the list $\{1, 2, \ldots, n\}$, its convex hull is called the is the permutahedron (it needs to be recentered) and $kxk_A$ is hard to compute (but can be used as a penalty).
Atomic norms

Suppose $k \cdot k_A$ is an atomic norm, focus on

$$
\begin{align*}
\text{minimize} & \quad k x k_A \\
\text{subject to} & \quad A x = A x_0
\end{align*}
$$

Proposition

**Optimality & recovery.** We write

$$
T_A(x_0) = \text{Cone}\{z - x_0 : k z k_A \leq k x_0 k_A\}
$$

the tangent cone at $x_0$. Then $x_0$ is the unique optimal solution of (3) iff

$$
T_A(x_0) \cap N(A) = \{0\}$$
Atomic norms

- Perfect recovery of $x_0$ by minimizing the atomic norm $kxk_A$ occurs when the intersection of the **subspace** $N(A)$ and the **cone** $T_A(x_0)$ is empty.

- When $A$ is i.i.d. Gaussian with variance $1/m$, the probability of the event $T_A(x_0) \cap N(A) = \{0\}$ can be bounded explicitly.

**Proposition**

**[Gordon, 1988]** Let $A \in \mathbb{R}^{m \times n}$, be i.i.d. Gaussian with $A_{i,j} \sim \mathcal{N}(0, 1/m)$, let

$\Omega = T_A(x_0) \cap B_2^p$ be the intersection of the cone $T_A(x_0)$ with the unit sphere, $x_0$ is the unique minimizer of (3) with probability $1 - \exp(-(\lambda_n - \omega(\Omega))^2/2)$ if

$$m \geq \omega(\Omega)^2 + 1$$

where

$$\omega(\Omega) = \mathbb{E} \left[ \sup_{y \in \Omega} y^T g \right] \quad \text{and} \quad \lambda_m = \frac{\sqrt{2} \Gamma((m+1)/2)}{\Gamma(m/2)}$$
Atomic norms

The previous result shows that computing the recovery threshold $n$ (number of samples required to reconstruct the signal $x_0$), it suffices to estimate

$$\omega(\Omega) = \mathbb{E} \left[ \sup_{y \in T_A(x_0), \|y\|_2 = 1} y^T g \right]$$

This quantity can be computed for many atomic norms $\| \cdot \|_A$.

- Suppose $x_0$ is a $k$-sparse vector, $k \times k_A = k \times k_1$ and
  $$\omega(\Omega)^2 \leq 2k \log(p/k) + 5k/4$$

- Suppose $X_0$ is a rank $r$ matrix in $\mathbb{R}^{m_1 \times m_2}$, then $kX \times k_A = kX \times k_*$ and
  $$\omega(\Omega)^2 \leq r(m_1 + m_2 - r)$$

- Suppose $X_0$ is an orthogonal matrix of dimension $n$, then $kX \times k_A = kX \times k_2$ and
  $$\omega(\Omega)^2 \leq \frac{3n^2 - n}{4}$$
References


