On the conditioning of random subdictionaries

by Joel A. Tropp

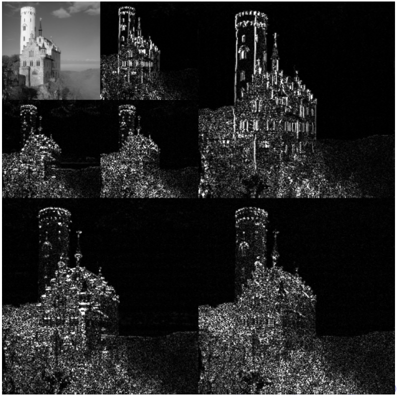
Leo Davy Martin Gjorgjevski

ENS Lyon
M2 Advanced Mathematics

November 2021

The goals of the presentation will be to understand

- The role of dictionaries for inverse problems
- The importance of conditioning and its relationship with sparsity
- How introducing randomness improves on the deterministic results



(sub)-dictionaries

Definition

A (sub)-dictionary Φ is a (sub collection of a) family of vectors, often called atoms, $\Phi = [\varphi_1, \cdots, \varphi_N]$.

Interest :

- Many possible choices, given a vector space a dictionary can be a basis, several basis, a frame,...
- Reconstructing signals
- Given a family of signals of interest it is possible to choose an adapted dictionary

Coherence and Spark

Signals can be written in $\mathbb{R}^d$, with d very large
Examples : sound record ($d \geq 10^4/s$), picture $d \geq 10^6$, video, ...
Is it possible to find the sparsest representation of a signal y with a given dictionary Φ ? i.e. solving

$$\min ||x||_0 : \quad y = \Phi x \tag{P0}$$

Concepts related to the solution of (P0) :

- Coherence of a dictionary $\mu = \sup_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$
- Spark of a dictionary= the largest number σ such that every set of σ columns of Φ is independent, $\sigma \geq \frac{1}{\mu}$
- Uncertainty principles

$\sqrt{d}$-bottleneck

Proposition 1

Any collection A of m columns of $\Phi = [\varphi_1, \cdots, \varphi_N]$ such that $||\varphi_i||_{\ell^2(\mathbb{R}^d)} = 1$ it holds that

$$||A^*A - I|| = \max\{\sigma_{\max}^2(A) - 1, 1 - \sigma_{\min}^2(A)\} \leq (m-1)\mu$$

Observing $\sigma_{\min}^2(A) > 0$ implies that the columns of A are linearly independent, then, if $(m-1)\mu < 1$

- Any solution of (P0) with less than m coefficients is the unique on its support
- Any solution of (P0) with less than $\frac{m}{2}$ is the unique sparsest solution

Difficulty : μ is bounded below by $\mathcal{O}(\frac{1}{\sqrt{d}})$ so we cannot say anything for sparsity constraint larger than $\mathcal{O}(\sqrt{d})$

(P0) and (P1) minimization problems

Solving (P0) :

$$\min_x ||x||_0 : \quad y = \Phi x$$

is very[1] hard in general.
Solving (P1):

$$\min_x ||x||_1 : \quad y = \Phi x$$

is much easier (linear programming).
In practice, (most of the time) solving P1 resolves P0 even if the solution has a sparsity constraint much larger than $\mathcal{O}(\sqrt{d})$.
Goals of the article :

- Show that (P0) is well posed with high probability with a sparsity constraint $\mathcal{O}(\frac{d}{\log d})$.
- Show the relationship between the sparsity constraint and the coherence μ

[1] NP Hard

From determinism to randomness

Deterministic : For all A made of m columns

- $||A^*A - I|| < 1 \implies$ unique sparse solution
- $||A^*A - I|| \leq C\mu m$
- $m \sim \mathcal{O}(\frac{1}{\mu}) \leq \mathcal{O}(\sqrt{d})$

Random : For A chosen uniformly among collections of m columns

- $\mathbb{E}||A^*A - I|| < 1$ or $\mathbb{P}(||A^*A - I|| \leq 1 - \delta)$ to get unique sparse solution with high probability
- $m \sim \mathcal{O}(\frac{d}{\log d})$

Probabilistic results

Theorem A

If Φ has $2d$ columns and $||\Phi x|| = 2||x||, \forall x$
X is a random m-columns dictionary, then

$$\mathbb{E}||X^*X - I|| \leq C\sqrt{\mu^2 m \log(m+1)}.$$

In particular, the solution of (P0) is unique if there exists a solution with less than $\frac{\mu^{-2}}{\log d}$ coefficients.

Theorem B

Let Φ dictionary with N columns and X a random m columns dictionary.

$$\text{If } \sqrt{\mu^2 m \log(m+1) s} + \frac{m}{N}||\Phi||^2 \leq c\delta \quad \text{with } s \geq 1$$

then, $\mathbb{P}(||X^*X - I|| \geq \delta) \leq m^{-s}$

How to apply these results ?

Theorem D

Let Φ a pair of o.n.b., X made of $|\Omega|$ columns from the first basis and $|T|$ from the second.
If $|\Omega| + |T| \leq \frac{c\mu^{-2}}{s \log d}$, then

$$\mathbb{P}(||X^*X - I|| \geq 0.5) \leq d^{-s}$$

If one has a model where signals are made with :

- arbitrary components in Ω
- random components in T

then on an event of large probability the sparsest solution is unique and the smaller the coherence the larger the support of the solution can be.

Compression of the hollow Gram matrix

How to prove the results ?

Restriction operator :

$$R_\Omega : \mathbb{C}^N \to \mathbb{C}^\Omega$$
$$(f_i)_{i=1}^N \mapsto (f_\omega)_{\omega\in\Omega}$$

Extension operator :

$$R_\Omega^* : \mathbb{C}^\Omega \to \mathbb{C}^N$$
$$(f_\omega)_\Omega \mapsto (f_\omega)_{\omega\in\Omega} + (0)_{[1,\cdots,N]\setminus\Omega}$$

We want to consider $X^*X - I$ where $suppX = \Omega$ which we can now rewrite as

$$X^*X - I = R_\Omega(\Phi^*\Phi - I)R_\Omega^* = R_\Omega H R_\Omega^*$$

Goal : Find (subgaussian) tail bounds of $||X^*X - I|| = ||R_\Omega H R_\Omega^*||$

Proposition 10

Let Z a non negative random variable such that $(\mathbb{E}Z^q)^{\frac{1}{q}} \leq \alpha\sqrt{q} + \beta, \forall q \geq Q$ then

$$\mathbb{P}(Z \geq e^{\frac{1}{4}}(\alpha u + \beta)) \leq e^{-\frac{u^2}{4}}, \quad \forall u \geq \sqrt{Q}$$

Theorem 9, Decoupling

Let R a restriction to m coordinates, A a hermitian matrix with $2N$ columns, T_1 and T_2 a partition of $\{1, \cdots, 2N\}$, then

$$(\mathbb{E}||RAR^*||^q)^{\frac{1}{q}} \leq 2 \max_{m_1+m_2=m} \left(\mathbb{E}||R_1 A_{T_1 \times T_2} R_2^*||^q\right)^{\frac{1}{q}}$$

where the R_i are independent restrictions to m_i coordinates of T_i.

The most important theorem of the article

With the results from the previous slide (and some work), obtaining a good bound on the moments of $||AR^*||$ would allow us to recover statements from the Theorems A, B and D.
The "most important theorem" of the article :

Theorem 8, Spectral norm of a random compression

A hermitian matrix with N columns, R restriction to m random coordinates.
Fix $q \geq 1$, then for each $p \geq \max\{2, 2\log(rank(AR^*), \frac{q}{2}\}$ it holds that

$$(\mathbb{E}||AR^*||^q)^{\frac{1}{q}} \leq 3\sqrt{p}||A||_{1,2} + \sqrt{\frac{m}{N}}||A||$$

where $||\cdot||_{1,2}$ is the maximum ℓ^2 norm of a column.

Theorem 8, Spectral norm of a random compression

$$(\mathbb{E}||AR^*||^q)^{\frac{1}{q}} \leq 3\sqrt{p}||A||_{1,2} + \sqrt{\frac{m}{N}}||A||$$

Corollary (Theorem C)

$\Phi = (\Phi_1, \Phi_2)$ *pair of o.n.b.,* Ω *arbitrary,* T *random,*
$m = \min(|\Omega|, |T|)$*, then*

$$\begin{aligned} \mathbb{E}||X^*X - I|| =& \mathbb{E}||(R_\Omega \Phi_1^* \Phi_2)R_T^*|| \\ & \leq C\sqrt{\mu^2|\Omega|\log(m+1)} + \sqrt{\frac{|T|}{d}} \end{aligned}$$

Proof.

$A = R_\Omega \Phi_1^* \Phi_2$, $q = 1$, $p = 2\log(m+1)$ since $rank(AR_T^*) \leq m$,
$||A||_{1,2} \leq \mu\sqrt{|\Omega|}$, $||A|| \leq 1$ □

Ideas behind the difficult theorems

Proving theorem 8 ?

- Banach space probability (Rudelson's lemma)
- Symmetrization
- Schatten norms
- Non-commutative Khintchine inequality

Proving theorem 9 ?

- Symmetrization again
- Probabilistic method

Proving proposition 10 ?

- Markov + moment bound
- Subgaussian concentration inequalities

Summary

- The problem (P0) is always well posed when the sparsity constraint is of order $\frac{1}{\mu} \leq \mathcal{O}(\sqrt{d})$
- Having small coherence μ (and therefore large spark σ) allows to recover large signals
- Introducing randomness allows (P0) to be well posed with high probability when the sparsity constraint is of order $\mathcal{O}(\frac{1}{\mu^2 \log(d)}) \leq \mathcal{O}(\frac{d}{\log d})$
- Applications to sparse recovery and sparse reconstruction