Semidefinite Programming with Applications in Geometry and Machine Learning.

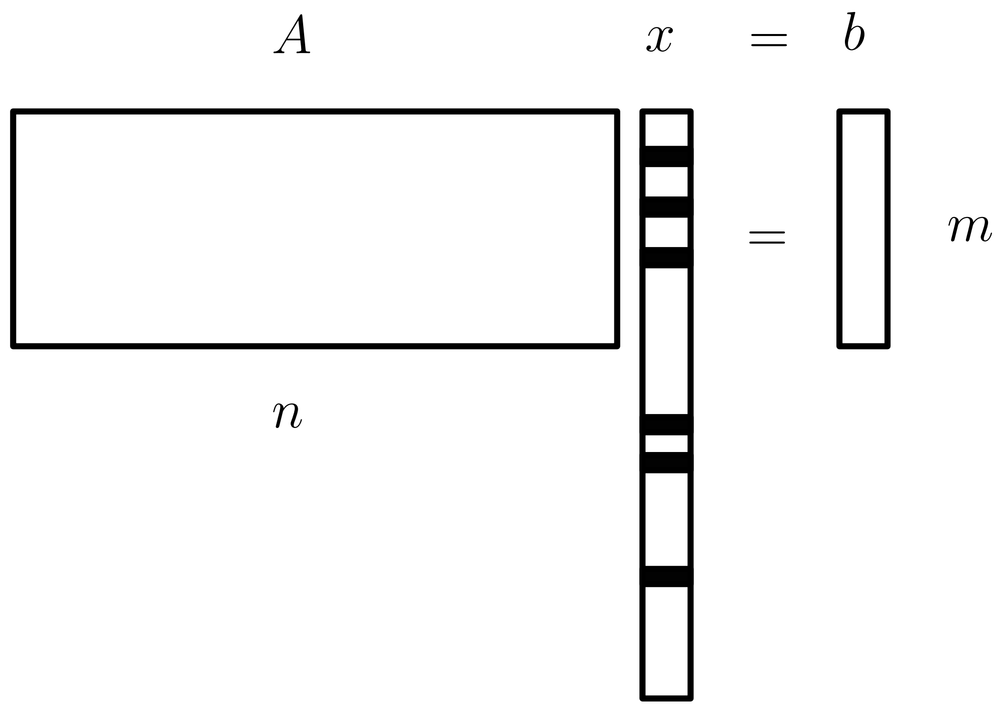
Part Two.

Alexandre d'Aspremont, *CNRS & Ecole Polytechnique*.

Support from NSF, ERC SIPA and Google.

Introduction

We seek to solve the following underdetermined linear system



where $A \in \mathbb{R}^{m \times n}$, with $n \geq m$, assuming the solution is **sparse**.

l_1 decoding

$$\begin{array}{ll} \text{minimize} & \mathbf{Card}(x) \\ \text{subject to} & Ax = Ae \end{array} \qquad \textbf{becomes} \qquad \begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = Ae \end{array}$$

- **Donoho and Tanner [2005], Candès and Tao [2005]:**

  *For some matrices A, when the solution e is sparse enough, the solution of the ℓ_1-**minimization** problem is also the **sparsest** solution to $Ax = Ae$.*

- This happens even when

$$\mathbf{Card(e) = O\left(\frac{m}{\log(n/m)}\right)}$$

when $m = \rho n$ and $n \to \infty$, which is provably optimal.

l_1 decoding

Many variants:

- The observations could be **noisy**.
- **Approximate solutions** might be sufficient.
- We might have strict **computational limits** on the decoding side.
- The **regression** setting has different objectives.

In this talk:

- Use the simplest linear **coding** problem formulation.
- Focus on the **complexity** of recovery conditions.

l_1 decoding: conditions

Conditions on the coding matrix A which guarantee recovery of all signals up to some cardinality k.

- **Incoherence:** bounds on the correlation between measurements

$$\mu(A) = \max_{i<j} |A_i^T A_j|$$

- **Nullspace property:** there is some $\alpha_k < 1/2$ such that

$$\|x\|_{k,1} \leq \alpha_k \|x\|_1, \quad \text{for all } x \in \mathcal{N}(A)$$

- **Restricted Isometry:** Let F s.t. $AF = 0$ and $\delta_k(F) = \max\{\delta_k^{\min}, \delta_k^{\max}\}$ with

$$(1 \pm \delta_k^{\max/\min}) = \begin{array}{ll} \text{max./min.} & x^T(FF^T)x \\ \text{s.t.} & \mathbf{Card}(x) \leq k \\ & \|x\| = 1, \end{array}$$

- **Etc. . .** See e.g. tutorial by [Indyk, 2008] or paper by [Van De Geer and Bühlmann, 2009]

l_1 decoding: main objective

Produce a **score** to identify good coding matrices A?

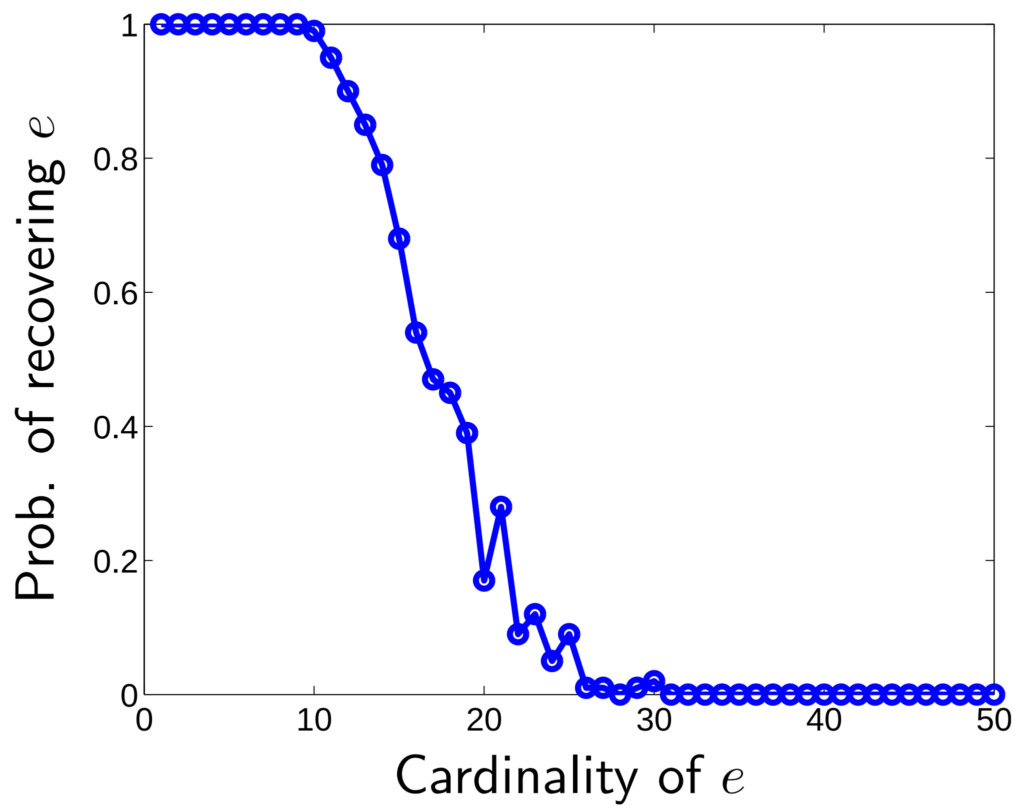
- **Ideally:** Given a matrix A, compute best threshold $k(A)$ such that exact l_1-decoding is guaranteed for all signals of cardinality up to $k(A)$.
- **In reality:** Exact thresholds are hard to compute. We would be happy with tractable scores which correlate with $k(A)$ but are easier evaluate.

l_1 decoding: main objective

Example: fix A, draw many random **sparse signals** e and plot the probability of perfectly recovering e when solving

$$\begin{array}{ll} \text{minimize} & \|x\|_1 \\ \text{subject to} & Ax = Ae \end{array}$$

in $x \in \mathbb{R}^n$ over 100 samples, with $n = 50$ and $m = 30$.



Motivation: dictionary learning

Consider the following **dictionary learning** problem [Mairal, Bach, Ponce, and Sapiro, 2009]. Given sample points $x_i \in \mathbb{R}^m$, solve

$$\min_{D \in \mathcal{C}} \sum_i \ell(x_i, D)$$

in the variable D, where the loss function is defined as

$$\ell(x_i, D) = \min_{\alpha} \|x_i - D\alpha\|_2^2 + \lambda \|\alpha\|_1$$

and $\mathcal{C}$ is some convex set. Mostly in a **compression** context here.

- The $\|\alpha\|_1$ penalty, as a proxy for cardinality, seeks "good signals".
- Usually, the set $\mathcal{C}$ is a norm ball, e.g. a normalization constraint $\|D_i\|_2 \leq 1$, which allows to identify D and α.

This is **learning without penalization**, i.e. potentially low generalization power.

How do we efficiently characterize good dictionaries?

l_1 decoding: conditions

A long wish list. . . Ideally, dictionary metrics should have the following features.

- **Universality:** prove reconstruction for all signals (or at least most signals).
- **Invariance:** recovery is a property of the nullspace only.
- **Low complexity:** tested in polynomial-time.
- **Error bound:** bound the decoding error.

l_1 decoding conditions: complexity

Conditions on the coding matrix A which guarantee recovery of all signals up to some cardinality k.

- **Incoherence:** Not universal, not invariant, easy to test but only guarantees recovery of signals of size $O(\sqrt{k^*})$ when the best performance is $O(k^*)$.
- **Restricted Isometry:** Universal, invariant. Also **hard to test:** the relaxation in d'Aspremont et al. [2007] shows recovery at cardinality $k = O(\sqrt{k^*})$ when A satisfies RIP at the threshold k^*. It provably cannot do better than that.
- **Nullspace property:** Universal, invariant. **Hard to test:** relaxations in d'Aspremont and El Ghaoui [2011], Juditsky and Nemirovski [2011] can prove exact recovery at cardinality $k = O(\sqrt{k^*})$ when A satisfies RIP at the threshold k^*. They provably cannot do better than that.

Outline

- Introduction
- **Geometrical conditions**
- Bounding the diameter

Geometrical conditions

Diameter

Kashin and Temlyakov [2007]: Very simple relationship between diameter of a section by A of the ℓ_1 ball and the recovery threshold k (largest signal size for which perfect recovery holds).

Proposition 2

Diameter & Recovery threshold. *Given a coding matrix* $A \in \mathbb{R}^{m\times n}$*, we write* x^{LP} *the solution of the* ℓ_1*-minimization LP and* e *the true signal. Suppose that there is some* $k > 0$ *such that*

$$\mathbf{diam}(B_1^n \cap \mathcal{N}(A)) = \sup_{\substack{Ax=0\\ \|x\|_1\leq 1}} \|x\|_2 \leq \frac{1}{\sqrt{k}} \qquad (1)$$

then sparse recovery $x^{\mathrm{LP}} = e$ *is guaranteed if* $\mathbf{Card}(e) < k/4$*, and*

$$\|e - x^{\mathrm{LP}}\|_1 \leq \; 4 \min_{\{\mathbf{Card}(y)\leq k/16\}} \|e-y\|_1.$$

Diameter

Proof. Kashin and Temlyakov [2007]. Suppose

$$\sup_{\substack{Ax=0\\ \|x\|_1\leq 1}} \|x\|_2 \leq k^{-1/2}$$

If x satisfies $Ax=0$, for any support set Λ with $|\Lambda| < k/4$,

$$\sum_{i\in\Lambda} x_i \;\leq\; \sum_{i\in\Lambda}|x_i| \;\leq\; \sqrt{|\Lambda|}\;\|x\|_2 \;\leq\; \sqrt{|\Lambda|/k}\;\|x\|_1 \;<\; \|x\|_1/2,$$

Let u be the true signal, with $\mathbf{Card}(u) < k/4$ and $\Lambda = \mathrm{supp}(u)$ 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 $x^{\mathrm{LP}} = u$. The error bound follows from similar arg.

Kashin decompostion

Results giving bounds on the diameter of random sections of the ℓ_1-ball can be traced back to Dvoretzky's theorem and the Kashin decomposition.

- **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.

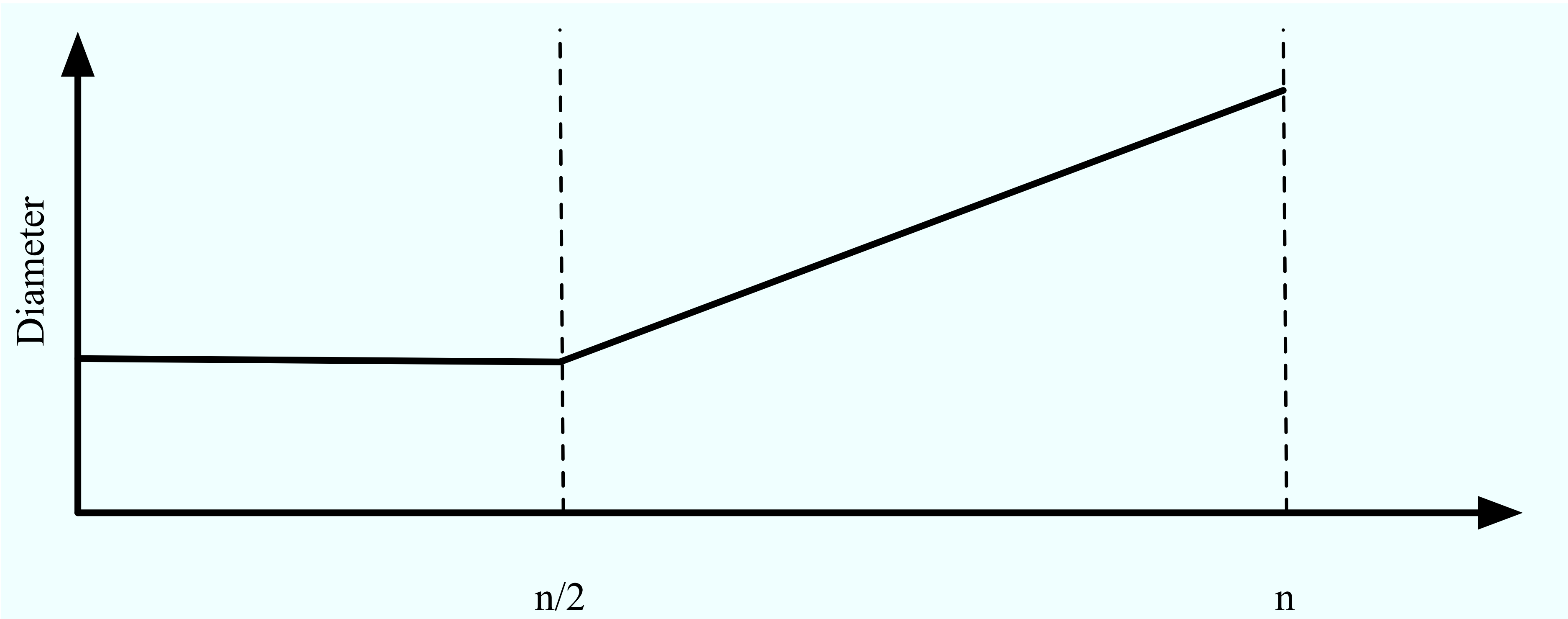
- For these subspaces, we have

$$\mathbf{diam}(B_1^n \cap E_i) \leq \frac{8}{\sqrt{n}}, \quad i = 1, 2,$$

  and we can guarantee ℓ_1 recovery of **all signals up to cardinality $n/64$** if we use a coding matrix with nullspace E_i.

Diameter & Random Sections

Schematically...



The diameter $\mathbf{diam}(B_1^n \cap E)$ decreases w.h.p. for smaller random sections, until these sections become almost spherical after which it does not change.

Diameter, low M^* estimate

Theorem 3

Low $\mathbf{M}^*$ estimate. *Let K be a symmetric convex body and $E \subset \mathbb{R}^n$ be a subspace of codimension k chosen uniformly w.r.t. to the Haar measure on $\mathcal{G}_{n,n-k}$, then*

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

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

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

ℓ_1-decoding: We have $M(B_\infty^n) \sim \sqrt{\log n / n}$ asymptotically. This means that random sections of the ℓ_1 ball with dimension $n - k$ have diameter bounded by

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

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

Deterministic Bounds on the Diameter

Bounding the diameter

Can we efficiently approximate the diameter of a **given** section of the ℓ_1 ball?

- Lovasz and Simonovits [1992] show that if we only have access to an oracle for a convex body K, then there is no randomized polynomial time algorithm to approximate the diameter of K within a factor $n^{1/4}$.
- Here however, we have much more information on the set K than a simple oracle. We know that
$$K = \{B_1^n \cap \mathcal{N}(A)\}.$$
The complexity of computing or approximating the diameter of such a set is unknown.

Bounding the diameter

Simple SDP relaxation: to bound

$$\mathbf{diam}(B_1^n \cap \mathcal{N}(A)) = \sup_{\substack{Ax=0\\ \|x\|_1 \leq 1}} \|x\|_2,$$

given a coding matrix A, we solve

$$SDP(A) \triangleq \max_{\substack{\mathbf{Tr}(A^TAX)=0\\ \|X\|_1 \leq 1,\, X \succeq 0}} \mathbf{Tr}\, X$$

which is a semidefinite program in $X \in \mathbf{S}_n$ (this is the classical lifting procedure where have have set $X = xx^T$). By construction

$$\mathbf{diam}(B_1^n \cap \mathcal{N}(A))^2 \leq SDP(A).$$

Bounding the diameter

Proposition 4

Relaxation performance. *Suppose* $A \in \mathbb{R}^{m \times n}$ *satisfies* $\mathbf{diam}(K \cap E) \leq 1/\sqrt{k}$ *the semidefinite relaxation will satisfy*

$$\sqrt{SDP(A)} \leq k^{-\frac{1}{4}}$$

Suppose now that n=2m, then we also have $(2n)^{-1/4} \leq \sqrt{SDP(A)}$ *and the semidefinite relaxation will certify exact decoding of all signals of cardinality at most* $O(\sqrt{m})$.

These results mean that the SDP relaxation will certify recovery at the threshold $\sqrt{k}$ when the true threshold is k, it cannot do better than that.

Estimating M^*

The low-M^* bound shows that we can use M^* as a good proxy for the diameter. . .

- We can apply the low-M^* bound in the **normed space** $\left\{\mathbb{R}^{n-k}, \|Fy\|_1\right\}$, where $AF = 0$, instead of the original normed space $\{\mathbb{R}^n, \|y\|_1\}$.
- Approximating $M^*(\{\|Fy\|_1 \leq 1\})$ simply means solving a lot of LPs.
- A section of a section is a section: taking random sections of this norm ball simply means **adding a few random rows** to the matrix A.
- From a compressed sensing point of view, $M^*(\{\|Fy\|_1 \leq 1\})$ simply measures how many additional experiments it would take to reach a given recovery performance with high probability.

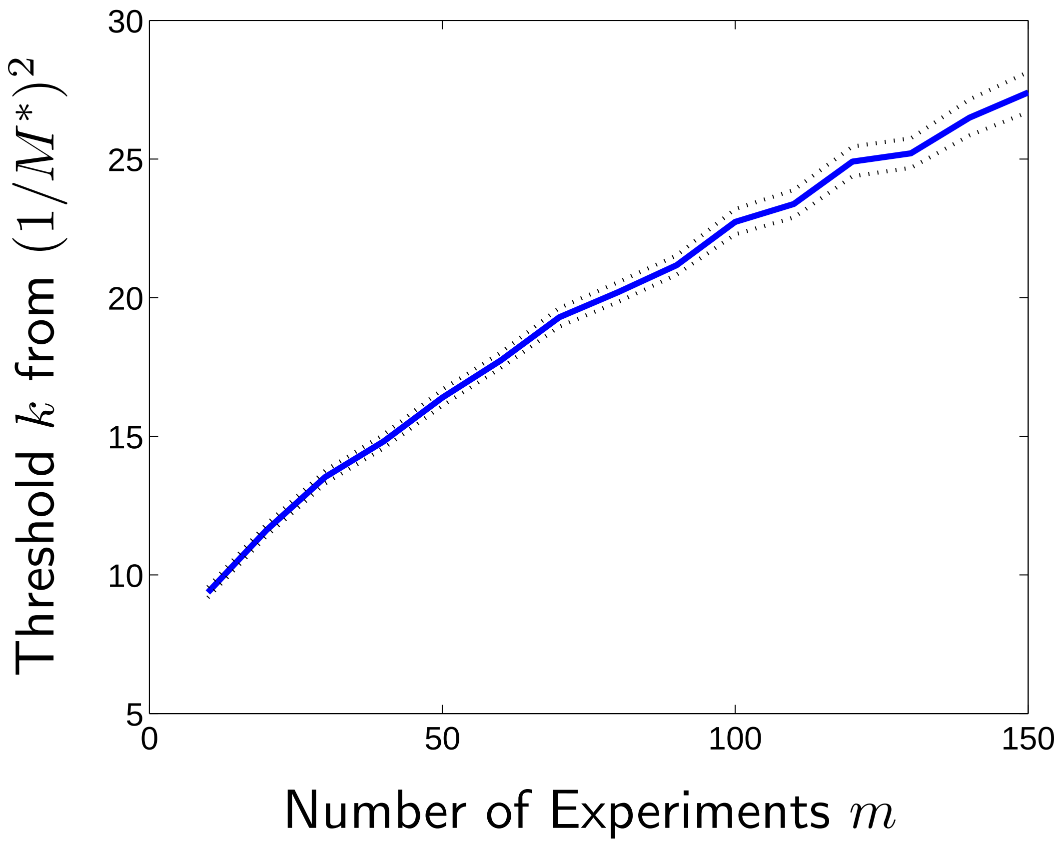
Estimating M^*

- We can estimate $M(K^*)$ by **simulation** [Bourgain et al., 1988, Giannopoulos and Milman, 1997, Giannopoulos et al., 2005]: if $K \subset \mathbb{R}^n$ is a symmetric convex body, $0 < \delta, \beta < 1$ and we pick N points x_i uniformly at random on the sphere $\mathbb{S}^{n-1}$ with
$$N = \frac{c \log(2/\beta)}{\delta^2} + 1$$
where c is an absolute constant, then
$$\left| M(K^*) - \frac{1}{N} \sum_{i=1}^{N} \|x_i\|_{K^*} \right| \leq \delta M(K^*)$$
with probability $1 - \beta$. Each sample requires solving a **linear program**.
- With high probability, we get a bound on the coding performance of the "enhanced" matrix A (A plus a few random measurements).

Estimating M^*

For good CS matrices, the bound $(1/M^*)^2$, which roughly controls the sparse recovery threshold through the low M^* estimate, should grow almost **linearly** with m

We estimate M^* for Gaussian sections of the ℓ_1 ball in $\mathbb{R}^{200}$, averaging 250 samples for each m and plot $(1/M^*)^2$ (dotted lines at 95% confidence).



Conclusion

- Increasingly large list of quality metrics for linear codes/dictionaries.
- Outside of coherence, most appear to be hard to approximate.
- Randomized polynomial time algorithm for testing the performance of slightly "enhanced" matrices.
- Direct connection with classical approximation problems.

Some open problems. . .

- Diameter and width are NP-Hard to approximate in the **oracle model**, but we have more structural information here. . .
- Can we derive **deterministic bounds** on M^* instead?
- Low M estimates also give bounds on the diameter. Estimating the **Dvoretzky dimension** for sections of the ℓ_1 ball is equivalent to solving a MAXCUT like problem. The $\pi/2$ approximation bound is insufficient here, can we do better?
- Use stochastic optimization algorithms for dictionary learning?

References

J. Bourgain, J. Lindenstrauss, and V. Milman. Minkowski sums and symmetrizations. *Geometric aspects of functional analysis*, pages 44–66, 1988.

E. J. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, 51(12):4203–4215, 2005.

A. d'Aspremont, L. El Ghaoui, M.I. Jordan, and G. R. G. Lanckriet. A direct formulation for sparse PCA using semidefinite programming. *SIAM Review*, 49(3):434–448, 2007.

Alexandre d'Aspremont and Laurent El Ghaoui. Testing the nullspace property using semidefinite programming. *Mathematical Programming*, 127:123–144, 2011.

D. L. Donoho and J. Tanner. Sparse nonnegative solutions of underdetermined linear equations by linear programming. *Proc. of the National Academy of Sciences*, 102(27):9446–9451, 2005.

A. Giannopoulos, V.D. Milman, and A. Tsolomitis. Asymptotic formulas for the diameter of sections of symmetric convex bodies. *Journal of Functional Analysis*, 223(1):86–108, 2005.

A. A. Giannopoulos and V. D. Milman. On the diameter of proportional sections of a symmetric convex body. *International Math. Research Notices, No. 1 (1997) 5–19.*, (1):5–19, 1997.

P. Indyk. Tutorial on compressed sensing (or compressive sampling or linear sketching). In *Workshop on Geometry and Algorithms*. Princeton, 2008.

A. Juditsky and A.S. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *Mathematical Programming Series B*, 127(57-88), 2011.

B. Kashin. The widths of certain finite dimensional sets and classes of smooth functions. *Izv. Akad. Nauk SSSR Ser. Mat*, 41(2):334–351, 1977.

B.S. Kashin and V.N. Temlyakov. A remark on compressed sensing. *Mathematical notes*, 82(5):748–755, 2007.

L. Lovasz and M. Simonovits. On the randomized complexity of volume and diameter. In *Foundations of Computer Science, 1992. Proceedings., 33rd Annual Symposium on*, pages 482–492. IEEE, 1992.

J. Mairal, F. Bach, J. Ponce, and G. Sapiro. Online dictionary learning for sparse coding. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 689–696. ACM, 2009.

A.S. Nemirovski. *Computation of matrix norms with applications to Robust Optimization*. PhD thesis, Technion, 2005.

Y. Nesterov. *Global quadratic optimization via conic relaxation*. Number 9860. CORE Discussion Paper, 1998.

A. Pajor and N. Tomczak-Jaegermann. Subspaces of small codimension of finite-dimensional banach spaces. *Proceedings of the American Mathematical Society*, 97(4):637–642, 1986.

S.A. Van De Geer and P. Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3: 1360–1392, 2009.