Weak Recovery Conditions using Graph Partitioning Bounds

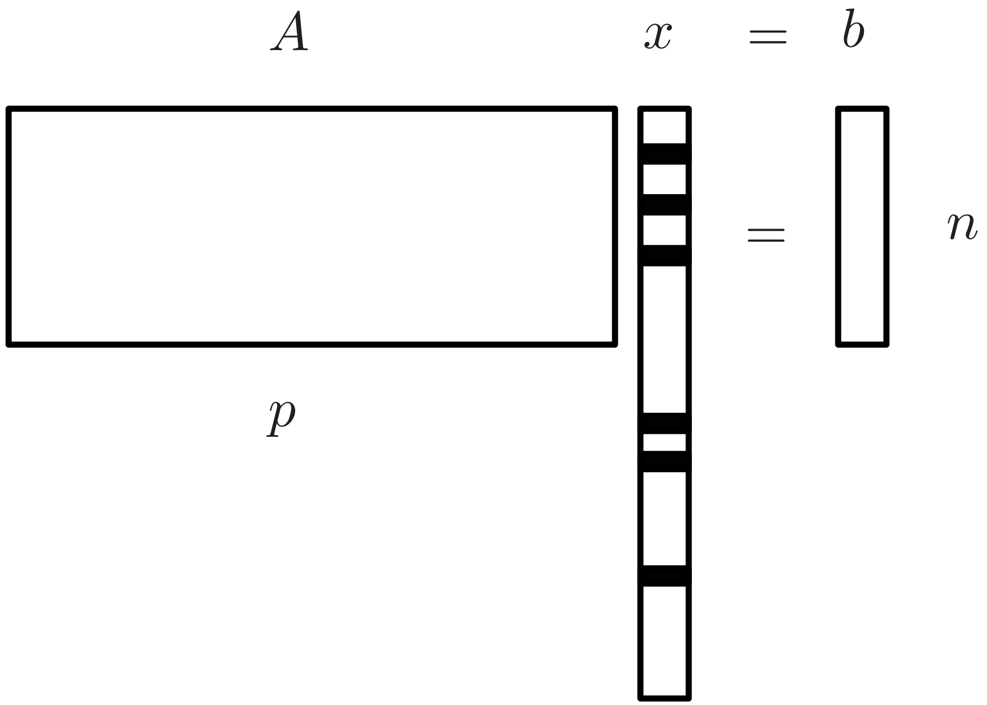
Alexandre d'Aspremont, *Princeton University*

Joint work with **Noureddine El Karoui**, *U.C. Berkeley*.

Support from NSF, DHS and Google.

Introduction

Consider the following underdetermined linear system



where $A \in \mathbf{R}^{n \times p}$, with $p \geq n$.

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?
  (Donoho, 2004; Donoho and Tanner, 2005; Donoho, 2006)

- **Coding:** Suppose we transmit a message which is corrupted by a few errors. How many errors does it take to start losing the signal?
  (Candès and Tao, 2005, 2006)

- **Statistics:** Variable selection & regression (LASSO, . . .).
  (Zhao and Yu, 2006; Meinshausen and Yu, 2008; Meinshausen et al., 2007; Candes and Tao, 2007; Bickel et al., 2007)

Simplification: the observations could be **noisy**, an **approximate solutions** might be sufficient, we might have strict **computational limits** on the decoding side.

l_1 relaxation

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

- **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

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

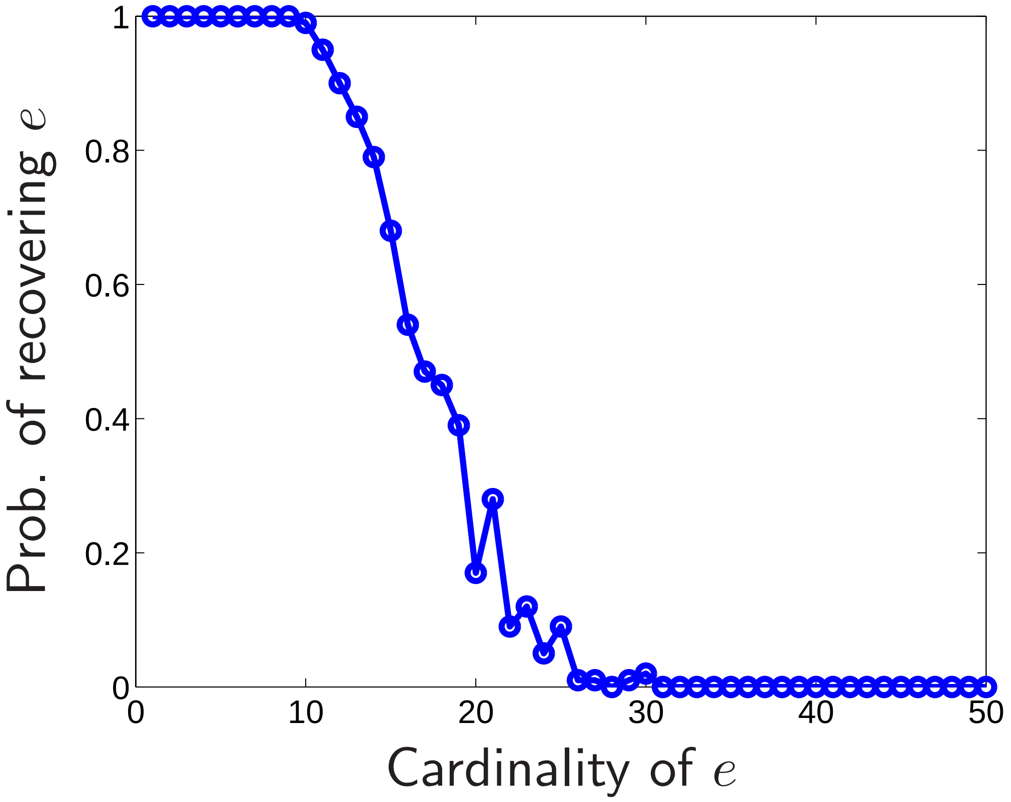
  asymptotically in p when $n = \rho p$, which is provably optimal.

Introduction

Illustration: 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 $x \in \mathbf{R}^p$ over 100 sample signals, with $p = 50$ and $n = 30$.



Introduction

Explicit conditions on the matrix A for perfect recovery of all sufficiently sparse signals e.

- **Nullspace Property (NSP):** Donoho and Huo (2001), Cohen et al. (2009).
- **Restricted Isometry Property (RIP):** Candès and Tao (2005).

Candès and Tao (2005) and Baraniuk et al. (2008) show that these conditions are satisfied by certain classes of **random matrices**: Gaussian, Bernoulli, etc. for near optimal values of $\mathbf{Card}(e)$. Donoho and Tanner (2005) used a geometric argument to obtain similar results.

Nullspace Property (NSP)

Given $A \in \mathbf{R}^{n \times p}$ and $k > 0$, Donoho and Huo (2001) or Cohen et al. (2009) among others, define the **Nullspace Property** of the matrix A as

$$\|x\|_{k,1} \leq \alpha_k \|x\|_1$$

for all vectors $x \in \mathbf{R}^p$ with $Ax = 0$, for some $\alpha_k \in [0, 1)$. Here $\|x\|_{k,1}$ is the ℓ_1 norm of the k largest (magnitude) coefficients in x.

Good CS matrices: nullspace populated with incoherent vectors.

Two thresholds:

- $\alpha_{2k} < 1$ means recovery is guaranteed by solving a ℓ_0 minimization problem.
- $\alpha_k < 1/2$ means recovery is guaranteed by solving a ℓ_1 minimization problem.

Nullspace Property (NSP)

The nullspace property constant **controls reconstruction error** when exact recovery does not occur. Suppose that there is some $\alpha_k < 1/2$ such that

$$\|x\|_{k,1} \leq \alpha_k \|x\|_1$$

for all $x \in \mathbf{R}^p$ with $Ax = 0$, then

$$\|x^{\mathrm{lp}} - e\|_1 \leq \frac{2}{(1 - 2\alpha_k)}\, r_k(e).$$

Here

$$r_k(e) = \min_{\mathbf{Card}(u) \leq k} \|u - e\|_1$$

is the **best possible approximation error**.

Restricted Isometry Property (RIP)

- Given $0 < k \leq p$, Candès and Tao (2005) define the **restricted isometry constant** $\delta_k(A)$ from **sparse eigenvalue** problems

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

in $x \in \mathbf{R}^p$, with $\delta_k(A) = \max\{\delta_k^{\min}, \delta_k^{\max}\}$.

- If $\delta_{2k}(A) < \sqrt{2} - 1$, we can recover the vector e **exactly** by solving

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

in the variable $x \in \mathbf{R}^p$. Here also, $\delta_{2k}(A)$ **controls reconstruction error** when exact recovery does not occur, with

Limits of performance

One small problem. . . Testing these conditions on general matrices is **harder** than finding the sparsest solution to an underdetermined linear system for example.

- SDP relaxation in d'Aspremont and El Ghaoui (2008) can prove exact recovery at cardinality $k = O(\sqrt{k^*})$ when A satisfies RIP at the threshold k^*. It cannot do better than $k = O(\sqrt{k^*})$.
- LP relaxation in Juditsky and Nemirovski (2008) guarantees the same $k = O(\sqrt{k^*})$ when A satisfies RIP at k^*. It cannot do better than this.
- The SDP relaxation for NSP in d'Aspremont et al. (2007) also fails beyond this threshold $k = O(\sqrt{k^*})$.

This means that all current convex relaxations for testing sparse recovery conditions (with known perf. bounds) cannot prove recovery beyond $\mathbf{O}(\sqrt{\mathbf{k}^*})$ for matrices satisfying sparse recovery conditions up to signal cardinality k^*. . .

Weak recovery conditions

Requiring recovery conditions to hold for **all** vectors e is perhaps too conservative.

- In many applications, satisfying these conditions with **high probability**, assuming a reasonable model on the signal e, would be sufficient.
- Main objective: produce conditions that can be tested efficiently to produce a **tractable measure of performance** for ℓ_1 recovery on **arbitrary matrices**.

Weak recovery conditions:

- Assume a distribution over e (ideally. . . we will take a shortcut here).
- Produce explicit conditions on the design matrix A for the NSP to hold with high probability, under this model.
- Derive tractable algorithms to check these conditions for values of the cardinality k much closer to the true threshold k^*.

Weak recovery conditions

Ideally. . .

- Start by defining a model for the **sparse** (or power law) signal e.
- Study the distribution of

$$x^{\mathrm{lp}} = \operatorname*{argmin}_{Ax=Ae} \|x\|_1$$

- Produce conditions on A for the NSP to hold with **high probability** on this distribution (the reconstruction errors $(x^{\mathrm{lp}} - e)$ are in the nullspace of A).

In practice here

- Extracting the distribution of $x^{\mathrm{lp}} - e$ from a model on e is hard (harder than the problem we are trying to solve).
- Instead, we directly **posit models on the nullspace**.
- Two different models: Gaussian or (rotated) bounded independent.
- Of course, these models could have zero measure w.r.t. the true model for the reconstruction error $x^{\mathrm{lp}} - e$.

Weak recovery conditions

Some severe shortcomings

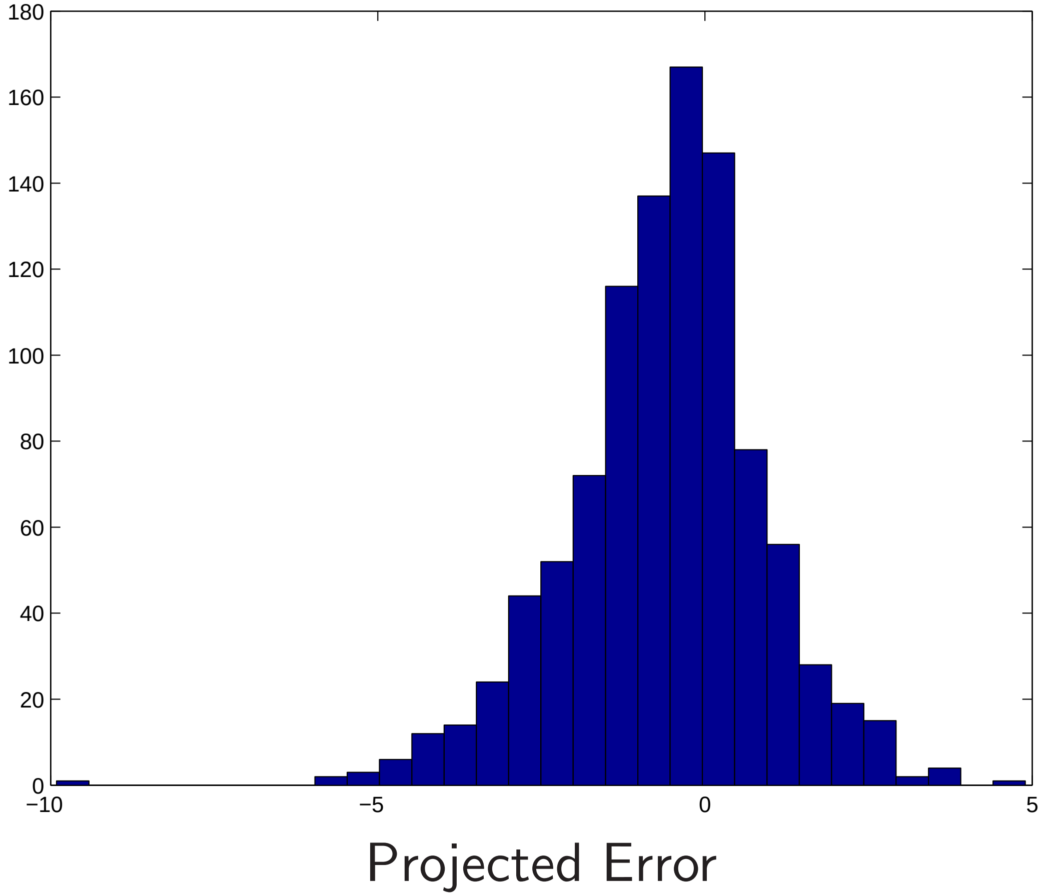
- We posit a model on the reconstruction error $(x^{\mathrm{lp}} - e)$ to test the NSP condition, which ultimately ensures that bounds on the norm $\|(x^{\mathrm{lp}} - e)\|_1$ hold. A bit wasteful at first sight. . .
- Favor tractability over statistical fidelity. Some empirical evidence that this is not completely off.

But a few interesting byproducts. . .

- Interesting link between concentration of norms and classic graph problems.
- These weak recovery conditions depend on good, **tractable** approximations.
- Cheap way of producing rough quantitative metrics on the quality of compressed sensing (i.e. design) matrices.

Only a thought experiment at this point. . .

Numerical results



Projected reconstruction error $v^T(x^{\mathrm{lp}} - e)$, along a fixed randomly chosen direction v, using a single Gaussian design matrix with $p = 100$, $n = 30$ and a thousand samples of a random sparse signal $e \in \mathbf{R}^{100}$ with 15 uniformly distributed coefficients.

Outline

- Introduction
- **Weak recovery conditions**
- Relaxation & approximation bounds
- Tightness & performance
- Numerical results

Weak recovery conditions: Gaussian model

Let us assume a **Gaussian model** Fy on the nullspace. Here $F \in \mathbf{R}^{p \times m}$ is a basis for the nullspace, so $AF = 0$. Can we check that

$$\|Fy\|_{k,1} \leq \alpha_k \|Fy\|_1$$

with high probability, when $y \sim \mathcal{N}(0, \mathbf{I}_m)$?

We can write $\|Fy\|_{k,1} = \max_{\{\|u\|_\infty \leq 1, \|u\|_1 \leq k\}} u^T Fy$ as a **max. of Gaussians**.

Concentration inequalities on Lipschitz functions of Gaussian variables then yield

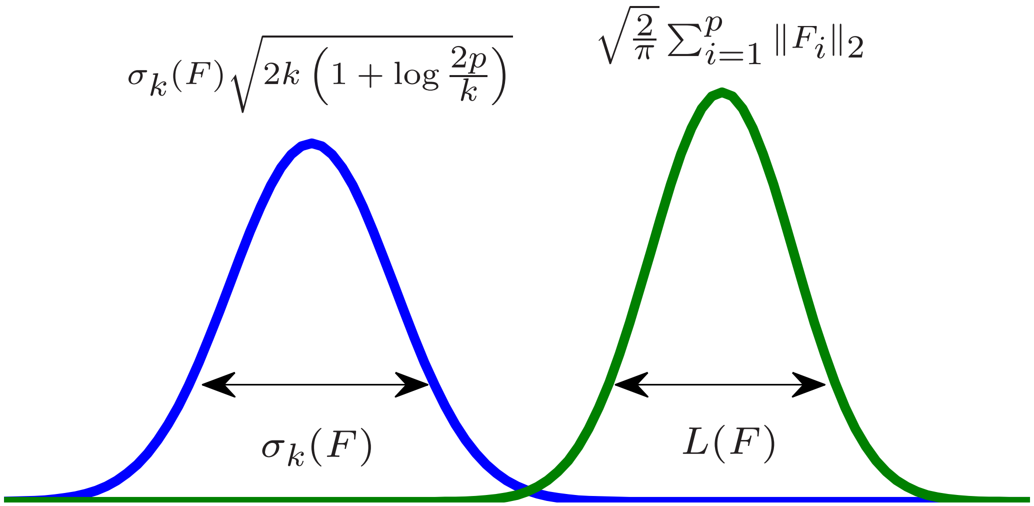
- $\textbf{Prob}\left[\|Fy\|_{k,1} \geq \left(\sqrt{2k\left(1 + \log \frac{2p}{k}\right)} + \beta\right)\sigma_k(F)\right] \leq e^{-\beta^2/2}$
- $\textbf{Prob}\left[\|Fy\|_1 \leq \left(\sqrt{2/\pi}\sum_{i=1}^p \|F_i\|_2 - \beta L(F)\right)\right] \leq e^{-\beta^2/2}$

where

$$\sigma_k^2(F) = \max_{\{u \in \{0,1\}^{2p}, \mathbf{1}^T u \leq k\}} u^T \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes FF^T u \quad \text{and} \quad L(F) = \sigma_p(F).$$

Bounding $\sigma_k(F)$ and $L(F)$

$$\|Fy\|_{k,1} \quad \text{and} \quad \|Fy\|_1$$


$\sigma_k(F)\sqrt{2k\left(1+\log\frac{2p}{k}\right)}$
$\sqrt{\frac{2}{\pi}}\sum_{i=1}^{p}\|F_i\|_2$
$\sigma_k(F)$
$L(F)$

In a Gaussian model:

- $\sigma_k(F)$ computed by k-**Dense-Subgraph**.
- $L(F)$ computed by **MAXCUT**.

Weak recovery conditions: Gaussian model

Here

$$\left(\sqrt{2k\left(1+\log\frac{2p}{k}\right)}+\beta\right)\sigma_k(F)\leq\left(\sqrt{\frac{2}{\pi}}\sum_{i=1}^{p}\|F_i\|_2-\beta L(F)\right)\alpha_k$$

ensures $\|Fy\|_{k,1}\leq\alpha_k\|Fy\|_1$ holds with high probability.

- Computing $\sigma_k(F)$ means solving a **k-Dense-Subgraph** problem

$$\sigma_k^2(F)=\max_{\{u\in\{0,1\}^{2p},\,\mathbf{1}^Tu\leq k\}}u^TMu,\qquad\text{with}\qquad M=\begin{pmatrix}1&-1\\-1&1\end{pmatrix}\otimes FF^T$$

- Computing $L(F)$ means solving a **MaxCut** type problem, directly related to the **MatrixCube** and **MatrixNorm** problems discussed in Nemirovski (2001) and Steinberg and Nemirovski (2005), or Ising spin glass models.

$$L^2(F)=\max_{v\in\{-1,1\}^p}v^TFF^Tv.$$

Weak recovery conditions: bounded model

Let us assume a **bounded model** on the nullspace. Let $F \in \mathbf{R}^{p \times m}$ be a basis for the nullspace, so $AF = 0$. Can we check that

$$\|Fy\|_{k,1} \leq \alpha_k \|Fy\|_1$$

with high probability, when the coefficients of y are bounded and independent? Note that F is defined up to a rotation, so we can assume some correlation in y.

$\|Fy\|_{k,1}$ is **Lipschitz, convex** in y so concentration inequalities yield

- $\mathbf{Prob}\left[\|Fy\|_{k,1} \geq \mathbf{E}[\|Fy\|_{1,k}] + \beta\sigma_k(F)\right] \leq e^{-\beta^2/2}$
- $\mathbf{Prob}\left[\|Fy\|_1 \leq \mathbf{E}[\|Fy\|_1] - \beta L(F)\right] \leq e^{-\beta^2/2}$

using the **same quantities $\sigma_k(F)$ and $L(F)$** as in the Gaussian model. The expectations can be computed efficiently by simulation.

Outline

- Introduction
- Weak recovery conditions
- **Relaxation & approximation bounds**
- Tightness & performance
- Numerical results

Bounding $\sigma_k(F)$ and $L(F)$

- A simple backward **greedy** algorithm produces a bound on $\sigma_k(F)$ tight up to a factor $(k/p)^2$.
- We can also bound $\sigma_k(F)$ using semidefinite relaxations, e.g.

$$SDP_k(M) = \begin{array}{ll} \text{max.} & \mathbf{Tr}\, MX \\ \text{s.t.} & 0 \leq X_{ij} \leq 1 \\ & \mathbf{Tr}\, X = k,\ X \succeq 0, \end{array}$$

  which is a semidefinite program in $X \in \mathbf{S}_p$.

- For $L(F)$, the classic **MaxCut** relaxation is tight up to a factor $2/\pi$, with

$$L^2(F) \leq \begin{array}{ll} \text{max.} & \mathbf{Tr}(XFF^T) \\ \text{s.t.} & \mathbf{diag}(X) = \mathbf{1}, X \succeq 0, \end{array}$$

  which is a semidefinite program in $X \in \mathbf{S}_p$.

Bounding $\sigma_k(F)$

Proposition 1

SDP tightness. *Suppose* $M \in \mathbf{S}_p$ *is positive semidefinite and* $k \geq p^{1/3}$. *Define*

$$\mathcal{D}_k(M) = \max_{\substack{u \in \{0,1\}^p \\ \mathbf{1}^T u \leq k}} u^T M u,$$

the relaxation

$$SDP_k(M) = \begin{array}{ll} \textit{max.} & \mathbf{Tr}\, MX \\ \textit{s.t.} & 0 \leq X_{ij} \leq 1 \\ & \mathbf{Tr}\, X = k,\ X \succeq 0, \end{array} \qquad (1)$$

satisfies

$$\frac{k}{p}\left(1 - \frac{o(1)}{k^{1/3}}\right)\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi} SDP_k(M)\right) \leq \mathcal{D}_k(M) \leq SDP_k(M),$$

where $G_{ij} = \sqrt{X_{ii}X_{jj}}$, $i, j = 1, \ldots, p$, *so in particular* $\mathbf{Tr}\, MG \geq 0$.

Bounding $\sigma_k(F)$

Approximation bounds (roughly match results on nonnegatively weighted graphs).

Proof (sketch). Hybrid randomization procedure, mixing sparse samples from Feige and Seltser (1997) and the argument of Nesterov (1998) on correlation. Generate points $w \in \{0,1\}^p$, with $w_i = u_i y_i$, where

$$u_i = \begin{cases} 1 & \text{with probability } q_i = k\frac{\sqrt{X_{ii}}}{\sum_{i=1}^p \sqrt{X_{ii}}}, \\ 0 & \text{otherwise.} \end{cases} \qquad \text{and} \qquad y_i = \begin{cases} 1 & \text{if } z_i \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

with $z \in \mathcal{N}(0, C)$ and $C_{ij} = X_{ij}/\sqrt{X_{ii}X_{jj}}$, $i, j = 1, \ldots n$. Then

$$\begin{aligned} \mathbf{E}[w^T M w] &= \frac{k^2}{S^2}\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi}\mathbf{Tr}(M(\arcsin(C) \circ G))\right) \\ &\geq \frac{k}{p}\left(\frac{1}{4}\mathbf{Tr}\, MG + \frac{1}{2\pi} SDP_k(M)\right) \end{aligned}$$

and $\mathbf{Prob}\left[\mathbf{Card}(u) \geq k\left(1 + k^{-1/3}\right)\right] \leq e^{-k^{1/3}/3}$.

Outline

- Introduction
- Weak recovery conditions
- Relaxation & approximation bounds
- **Tightness & performance**
- Numerical results

Weak NSP versus RIP

Computing the **RI constant** δ_k means solving

$$(1+\delta_k^{\max}) = \max_{\substack{u\in\{0,1\}^p\\ \mathbf{1}^T u\leq k}} \; \max_{\|x\|=1} \; u^T(FF^T\circ xx^T)u$$

in $x\in\mathbf{R}^p$, $u\in\{0,1\}^p$.

Computing the **weak NSP constant** $\sigma_k(F)$ defined above

$$\sigma_k(F) = \max_{\substack{u\in\{0,1\}^{2p}\\ \mathbf{1}^T u\leq k}} \; u^T M u$$

in $u\in\{0,1\}^{2p}$, where $M = \begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \otimes FF^T$.

Another interpretation: $\sigma_k(F)$ is a wider measure of **incoherence** (on submatrices of dimension k).

Limits of performance

Suppose the matrix $F^T \in \mathbf{R}^{m\times p}$ satisfies the **restricted isometry property** (RIP) with constant $\delta_k > 0$ at cardinality k, then

$$\sigma_k(F) \leq \sqrt{k(1+\delta_k)} \quad \text{and} \quad \|F_i\|_2 \geq \sqrt{1-\delta_1}$$

and $L(F) \leq p\sqrt{(1+\delta_k)/k}$.

In the Gaussian model, we can show that **weak NSP** is indeed **weaker than RIP** (and much easier to test).

Proposition 2

Weak recovery and RIP. *Let $n = \mu p$ and $k = \kappa n \log^{-1}(p/k)$ for some $\mu, \kappa \in (0,1)$. Suppose $F^T \in \mathbf{R}^{m\times p}$ satisfies the restricted isometry property with constant δ_k with $0 < \delta_k < c < 1$ at cardinality k, where c is an absolute constant, then F satisfies the weak recovery condition for p large enough.*

Limits of performance

Proposition 3

Tightness. *Suppose the matrix* $F \in \mathbf{R}^{p\times m}$ *satisfies the weak recovery condition up to cardinality* $k^* = \gamma(p)p$ *for some* $\gamma(p) \in (0,1)$, $\beta > 0$ *and* $\alpha_{k^*} \in [0,1]$, *i.e.*

$$\left(\sqrt{2k^* \log\frac{2p}{k^*}} + \beta\right) \sigma_{k^*}(F) \leq \left(\sqrt{\frac{2}{\pi}}\sum_{i=1}^{p} \|F_i\|_2 - \beta L(F)\right) \alpha_{k^*},$$

and let $SDP_k(\cdot)$ *be defined as in (1), we have*

$$\left(\sqrt{2k \log\frac{2p}{k}} + \beta\right) (SDP_k(M))^{1/2} \leq \left(\sqrt{\frac{2}{\pi}}\sum_{i=1}^{p} \|F_i\|_2 - \beta L(F)\right) \alpha_{k^*},$$

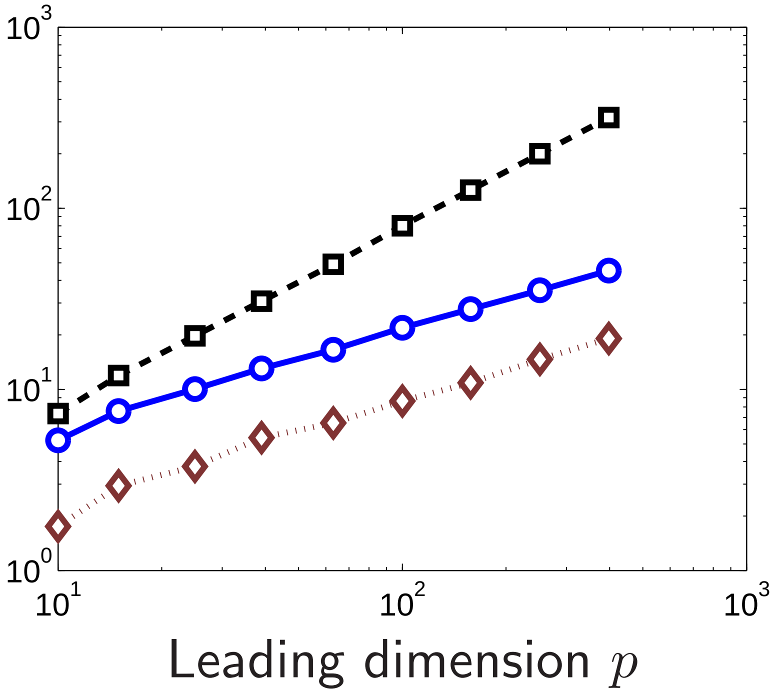
for p *sufficiently large, when* $k \leq \gamma(p)(\log p)^{-1}k^*$, *with* M.

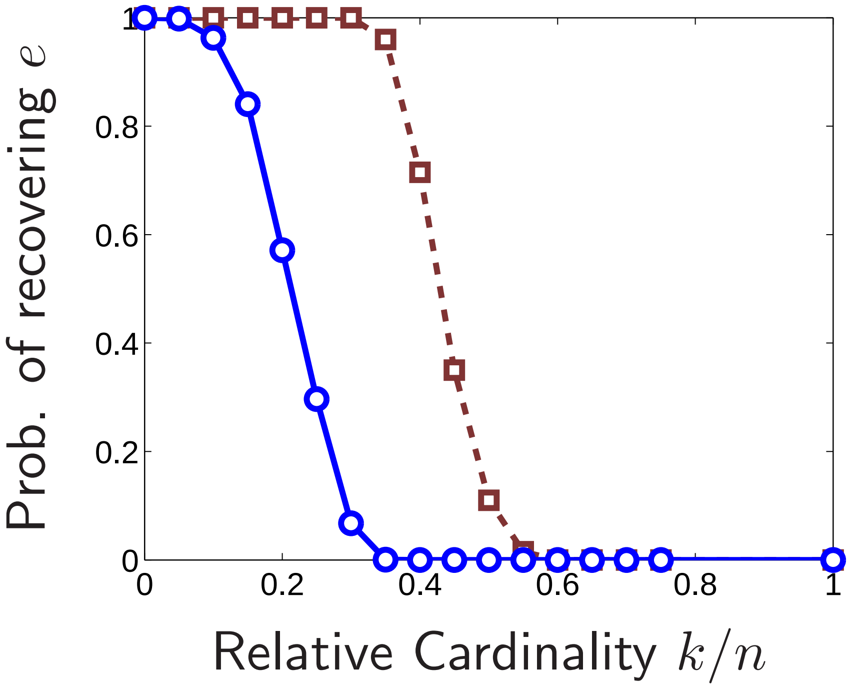
In other words, if F satisfies the weak recovery conditions at cardinality $k^* = \gamma(p)p$, the SDP relaxation will certify it up to $k = \gamma^2(p)(\log p)^{-1}p$.

Outline

- Introduction
- Weak recovery conditions
- Relaxation & approximation bounds
- Tightness & performance
- **Numerical results**

Numerical results





Left: Loglog plot of mean values of $L(F)$ (blue circles), $\sigma_k(F)$ (brown diamonds) and $\sum_{i=1}^{p} \|F_i\|_2$ (black squares) for Gaussian matrices of increasing dimensions p, with $m = p/2$.
Right: Predicted (blue circles) versus empirical (brown squares) probability of recovering the true signal e, where $F \in \mathbf{R}^{p \times n}$ is Gaussian with $n = p/2$, for various values of the relative cardinality k/n.

Conclusions

- Testing that the NSP holds with high probability seems to be much easier than checking that it always holds.
- When the design matrix satisfies RIP at the optimal regime where $\mathbf{Card}(e) = O\left(n/\log(p/n)\right)$, the corresponding weak conditions also hold.
- The constant α_k in the weak conditions provides a **rough but tractable measure of performance** for ℓ_1 recovery using arbitrary design matrices.

Some important questions unanswered here.

- Our model is defined on the reconstruction error. Ideally, we should have modeled the signal e directly.
- Even if that's not possible in the general case, can we at least calibrate a good model for $x^{\mathrm{lp}} - e$ using statistics on e?
- Better approximation bounds on sparse eigenvalues, NSP, $\sigma_k(F)$?

*

References

R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constructive Approximation*, 28(3):253–263, 2008.

P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *Preprint Submitted to the Annals of Statistics*, 2007.

E. Candes and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

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

E.J. Candès and T. Tao. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Transactions on Information Theory*, 52(12):5406–5425, 2006.

A. Cohen, W. Dahmen, and R. DeVore. Compressed sensing and best k-term approximation. *Journal of the AMS*, 22(1):211–231, 2009.

A. d'Aspremont and L. El Ghaoui. Testing the nullspace property using semidefinite programming. *To appear in Mathematical Programming*, 2008.

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.

D. L. Donoho. Neighborly polytopes and sparse solution of underdetermined linear equations. *Stanford dept. of statistics working paper*, 2004.

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.

D.L. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 52(4):1289–1306, 2006.

D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Transactions on Information Theory*, 47(7): 2845–2862, 2001.

U. Feige and M. Seltser. On the densest k-subgraph problem. Technical report, Department of Applied Mathematics and Computer Science, The Weizmann Institute, 1997.

A. Juditsky and A.S. Nemirovski. On verifiable sufficient conditions for sparse signal recovery via ℓ_1 minimization. *ArXiv:0809.2650*, 2008.

N. Meinshausen and B. Yu. Lasso-type recovery of sparse representations for high-dimensional data. *Annals of Statistics*, 37(1):246–270, 2008.

N. Meinshausen, G. Rocha, and B. Yu. A tale of three cousins: Lasso, l2boosting, and danzig. *Annals of Statistics*, 35(6):2373–2384, 2007.

A.S. Nemirovski. The matrix cube problem: Approximations and applications. *INFORMS*, 2001.

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

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

P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.