High-Dimensional Data Analysis

Lecture 10 - Convex Methods for Sparse Signal Recovery

Fall semester - 2024

Dr. Eng. Valentin Leplat

Innopolis University

November 7, 2024

Outline

Geometric Intuition

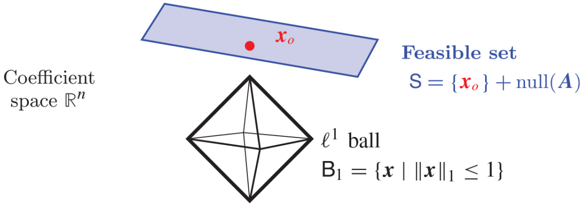
Geometric Intuition: Coefficient Space

Given $y = Ax_o \in \mathbb{R}^m$ with $x_o \in \mathbb{R}^n$ sparse:

$$\min_x \|x\|_1 \qquad \text{subject to } Ax = y \tag{1}$$

The space of all feasible solutions is an affine subspace:

$$S = \{x|Ax = y\} = \{x_o\} + \text{null}(A) \subset \mathbb{R}^n. \tag{2}$$

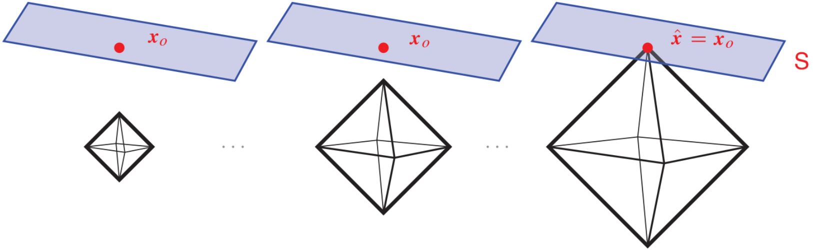


ℓ^1 Minimization in the Coefficient Space

Gradually expand a ℓ^1 ball of radius t from the origin 0:

$$t.B_1 = \{x \mid \|x\|_1 \leqslant t\} \subset \mathbb{R}^n, \tag{3}$$

till its boundary first touches the feasible set S:



Note: the ℓ^1 ball is "pointy" along the axes.
The ℓ^1 recovery problem is to pick out a point in S that has the minimum ℓ^1 norm. We can see that $\hat{x}$ is such a point.

Comparison between ℓ^1 and ℓ^2 Minimization

Given $y = Ax_o \in \mathbb{R}^m$ with $x_o \in \mathbb{R}^n$ sparse:

$$\begin{aligned} \mathbf{A}: \quad & \min_x \|x\|_1 \qquad \text{subject to } Ax = y. \\ & \text{versus} \\ \mathbf{B}: \quad & \min_x \|x\|_2 \qquad \text{subject to } Ax = y. \end{aligned} \tag{4}$$

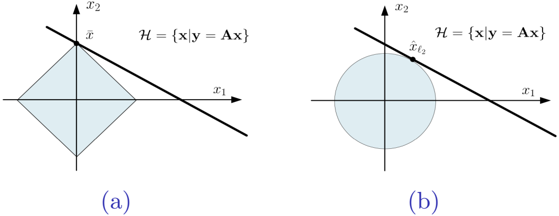


Figure: (a): shows the ℓ^1 recovery solution. The point $\bar{x}$ is a "sparse" vector; the line $\mathcal{H}(= S)$ is the set of all x that satisfy $y = Ax$. (b) shows the geometry when ℓ^2 norm is used. We can see that the solution $\hat{x}$ may not be sparse.

Sparsity Promoting with Different ℓ^p Norms

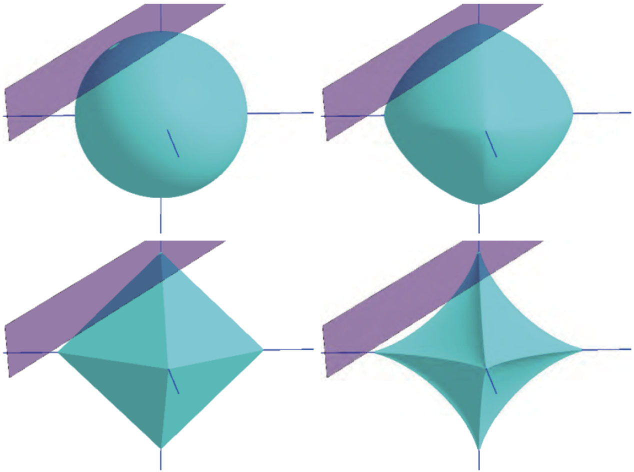


Figure: Intersection between the ℓ^p-ball and the feasible set S, for $p = 2$, 1.5, 1 and 0.7, respectively. (Some argue $p = 0.5$ is somewhat special.)

A First Correctness Result via Incoherence

Outline

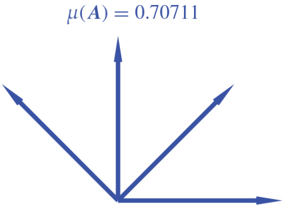
Coherence of a Matrix

Definition: Mutual Coherence

For a matrix $A = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \end{pmatrix} \in \mathbb{R}^{m \times n}$ with nonzero columns, the *mutual coherence* $\mu(A)$ is the largest normalized inner product between two distinct columns:

$$\mu(A) = \max_{i \neq j} \left| \left\langle \frac{a_i}{\|a_i\|_2}, \frac{a_j}{\|a_j\|_2} \right\rangle \right|. \tag{5}$$

Example:



Outline

Uniqueness of Sparse Solution

Proposition: Coherence controls Kruskal Rank

For any $A \in \mathbb{R}^{m \times n}$,

$$\operatorname{krank}(A) \geqslant \frac{1}{\mu(A)}. \tag{6}$$

In particular, if $y = Ax_o$ and

$$\|x_o\|_0 \leqslant \frac{1}{2\mu(A)}, \tag{7}$$

then x_o is the unique optimal solution to the ℓ^0 minimization problem

$$\min_x \|x\|_0 \qquad \text{s.t. } Ax = y \tag{8}$$

Proof: (more details on the board)

$$1 - k\mu(A) < \sigma_{\min}(A_l^H A_l) \leqslant \sigma_{\max}(A_l^H A_l) < 1 + k\mu(A)$$

Correctness of ℓ^1 Minimization

Theorem 1: ℓ^1 succeeds under incoherence

Let A be a matrix whose columns have unit ℓ^2 norm, and let $\mu(A)$ denote its mutual coherence. Suppose that $y = Ax_o$, with

$$\|x_o\|_0 \leqslant \frac{1}{2\mu(A)}. \tag{9}$$

Then x_o is the unique optimal solution to the problem

$$\min_x \|x\|_1 \qquad \text{s.t.} \;\; Ax = y \tag{10}$$

Correctness of ℓ^1 Minimization

Remarks:

- it is possible to improve the condition of the Theorem slightly, to allow recovery of x_o satisfying $\|x_o\|_0 \leqslant \frac{1}{2}(1 + \frac{1}{\mu(A)})$
- however this is tight !, indeed, this is the best possible statement, since there exist examples of A and x_o with $\|x_o\|_0 > \frac{1}{2}\left(\frac{1}{\mu(A)} + 1\right)$ for which ℓ^1 minimization does not recover x_o.
- Know that certain classes of A (Matrices with *Restricted Isometry Property*) of practical importance, far better guarantees are possible,
- and this has important implications for sensing, error correction, and number of related problems. [1]

[1] but that's beyond the scope of this humble introduction.

Correctness of ℓ^1 Minimization

Given $y = Ax_o$, try to find x_o via ℓ^1 Minimization:

$$\min_x \|x\|_1 \qquad \text{s.t. } Ax = y \tag{11}$$

Lagrangian formulation:

$$\min \|x\|_1 + \lambda^H(y - Ax), \qquad \exists \lambda \in \mathbb{R}^m \tag{12}$$

Optimality condition: x_o is the minimum of $f(x)$ if and only if 0 is in the subgradient $\partial f(x)$ at x_o:

$$f(x) \geqslant f(x_o) + \langle 0, x - x_o \rangle. \tag{13}$$

Optimality condition for ℓ^1 Minimization:

$$\begin{aligned} 0 &\in \partial\|x_o\|_1 - A^H\lambda \\ \equiv A^H\lambda &\in \partial\|x_o\|_1 \end{aligned} \tag{14}$$

Correctness of ℓ^1 Minimization

Proof (a sketch of key ideas):
Due to convexity of $\|.\|_1$, for any $v \in \partial\|.\|_1(x_o)$ and $x^{'} \in \mathbb{R}^n$(feasible),

$$\|x^{'}\|_1 \geqslant \|x_o\|_1 + \langle v, x^{'} - x_o \rangle \tag{15}$$

For $v = A^H\lambda$, we have $\langle A^H\lambda, x^{'} - x_o\rangle = \langle \lambda, A(x^{'} - x_o)\rangle = 0$. Therefore

$$\|x^{'}\|_1 \geqslant \|x_o\|_1 \tag{16}$$

To find such an optimality certificate $A^H\lambda \in \partial\|.\|_1(x_o)$, we need:

$$A_l^H\lambda = \sigma, \qquad \|A_{l^c}^H\lambda\|_\infty \leqslant 1 \tag{17}$$

A natural "candidate":

$$\hat{\lambda} := A_l(A_l^H A_l)^{-1}\sigma \tag{18}$$

The rest is to check this satisfies Equations (17) under the given conditions.

Correctness of ℓ^1 Minimization

By construction $A_l^H \hat{\lambda} = \sigma$. We are just left to verify second condition from (17), that it check if

$$\|A_{l^c}^H \hat{\lambda}\|_\infty = \|A_{l^c}^H A_l (A_l^H A_l)^{-1} \sigma\|_\infty \leqslant 1 \tag{19}$$

Without loss of generality, consider any single element of this vector ($j \in l^c$) which has the form:

$$\begin{aligned} |a_j^H A_l (A_l^H A_l)^{-1} \sigma| &\leqslant \underbrace{\|A_j^H a_j\|_2}_{\leqslant \sqrt{k}\mu(A)} \underbrace{\|(A_l^H A_l)^{-1}\|_2}_{< \frac{1}{1-k\mu(A)}} \underbrace{\|\sigma\|_2}_{=\sqrt{k}} \\ &< \frac{k\mu(A)}{1-k\mu(A)} \\ &\leqslant 1 \quad \text{provided } k\mu(A) \leqslant \frac{1}{2} \end{aligned} \tag{20}$$

This concludes the (sketch of) proof.

Outline

Constructing Incoherent Matrices

In previous theorem, we have seen that if $\|x_o\|_0 \leqslant 1/(2\mu(A))$, x_o is correctly recovered by ℓ^1 min. According to this result, matrices with smaller coherence admit better (higher) bounds.
Examples:

1. For two orthogonal matrices, say Φ is the classic Fourier Transform bases and Ψ is the identity I or certain wavelet transform bases,

$$A = \begin{pmatrix} \Phi & \Psi \end{pmatrix} \in \mathbb{C}^{n \times 2n}.$$

2. Another case which is of great interest is when the matrix A has the form

$$A = \Phi_l^H \Psi$$

   where $l \subset [n]$, and $\Phi_l \in \mathbb{R}^{n \times |l|}$ is a submatrix of an orthogonal base.
   For example, in the MRI problem in the previous chapter, Φ would correspond to the (Discrete) Fourier Transform, while Ψ was the basis of sparsity (e.g., wavelets).

Constructing Incoherent Matrices

As it turns out, incoherence is a generic property for almost all matrices.
So the easiest way to build a matrix A with small $\mu(A)$ is simply to choose matrix at random. The following theorem makes this precise:

Theorem 2

Let $A = [a_1 | \cdots | a_n]$ with columns $a_i \sim \text{uni}(\mathcal{S}^{m-1})$ chosen independently according to the uniform distribution on the sphere. Then with probability at least $3/4$,

$$\mu(A) \leqslant C\sqrt{\frac{\log(n)}{m}}$$

where $C > 0$ is a numerical constant.

Constructing Incoherent Matrices

Several points about the previous Theorem, there is nothing particularly special

- about the success probability 3/4: some tricks are possible to affect the constant C, and make the success probability arbitrarily close to 1.
- About the the uniform distribution on $\mathcal{S}^{m-1}$ - many distributions will produce similar results.

Outline

Limitations of Incoherence

- Theorem 1 gives a quantitative trade-off between niceness of A and sparsity of x_o, which asserts that when x_o is sparse enough, $\|x_o\|_0 \leqslant 1/(2\mu(A))$, then x_o is the unique optimal solution to the ℓ^1 min problem. [2]
- **How sharp is this result ?**
  According to Theorem 2, a random matrix $A \in \mathbb{R}^{m \times n}$ with high probability has its coherence bounded from above as $\mu(A) \leqslant C\sqrt{\log(n)/m}$.

[2] this gives a sufficient condition for the ℓ^1 minimization to be correct

Limitations of Incoherence

- So, for a "generic" A, the recovery guarantee implies correct recovery of x_o with $O(\sqrt{m/\log(n)})$ nonzeros !
- If we turn that around, and think of the matrix multiplication $x \to Ax$ as a sampling procedure, then for appropriately distributed random A, we can recover k-sparse x_o from

$$m \geqslant C^{'} k^2 \log(n)$$

  observations.
- When k is small, this is better than simply sampling all n entries of x....
- but the measurement burden $m = \Omega(k^2)$ seems a "little bit too high" when you think about it !
- **indeed**: to specify a k-sparse x, we only need to specify its k non-zero entries,.., and yet the theory demands k^2 samples !!

Limitations of Incoherence

- One might naturally guess that the choice of A as a random matrix was a poor one..
- ... perhaps some *delicate deterministic* construction can yield a better performance guarantee, by making $\mu(A)$ smaller.?
- As it turns out in this case, no matter what we do, we cannot construct a matrix whose coherence is significantly smaller than a randomly chosen one; this is the consequence of the lower-bound for the coherence $\mu(A)$ as demonstrated by Welch ("Welch bound").

Theorem 3: Welch Bound

For any matrix $A = [a_1|\cdots|a_n] \in \mathbb{R}^{m\times n}$, $m \leqslant n$, suppose that the columns a_i have unit ℓ^2 norm. Then

$$\mu(A) \geqslant \sqrt{\frac{n-m}{m(n-1)}}$$

Limitations of Incoherence

The important thing to notice: if we take n proportional to m, i.e. $n = \beta m$ for some $\beta > 1$, then the bound says that for *any* $m \times n$ matrix A,

$$\mu(A) \geqslant \Omega\left(\frac{1}{\sqrt{m}}\right).$$

Hence, in the *best possible case*, Theorem 1 guarantees we can recover x_o with about $\sqrt{m}$ nonzero entries → demand $m \geqslant C'' k^2$ samples.

Limitations of Incoherence

How does the coherence decay with dimension ?:

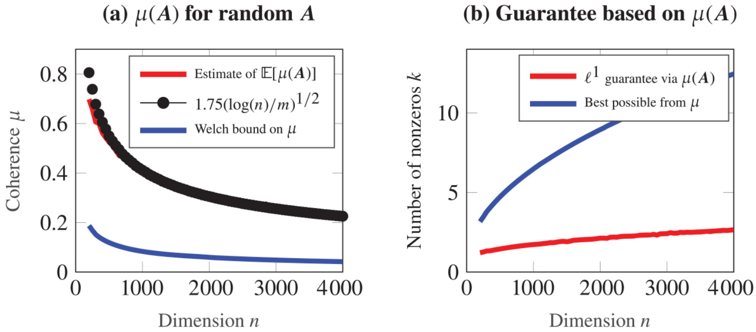


Figure: (a) - Average $\mu(A)$ over 50 trials, with $A \sim_{\text{iid}} \text{uni}(\mathcal{S}^{m-1})$ for various n and $m = n/8$. Black curve is for reference, and blue curve is the Welch bound (the min achievable $\mu(A)$).
(b) - Average number of non-zeros k which we can guarantee to reconstruct using the observed $\mu(A)$ and Theorem 1 (red curve). The blue curve bounds the best possible number of non-zeros entries using Theorem 1, for *any* matrix $A \in \mathbb{R}^{m\times n}$ using the Welch Bound.

Limitations of Incoherence

Incoherence ensures to recover k-sparse solution from

$$m \geqslant O(k^2)$$

measurements.
Experimental results suggest $m = O(k)$:
In a proportional growth setting $m \propto n$, $k \propto m$, ℓ^1 *minimization succeeds with very high probability whenever the constants of proportionality* n/m *and* k/m *are small enough.*

How to sharpen the bound?

A: need a more refined measure of goodness of A than the rather crude coherence or incoherence, out of the scope.

Convex Methods for Low-Rank Matrix Recovery

Outline

Problem

Recovering a sparse signal x_o:

$$\underbrace{y}_{\textit{observations}} = A \underbrace{x_o}_{\textit{unknown}} \tag{21}$$

where $A \in \mathbb{R}^{m \times n}$ is a linear map.

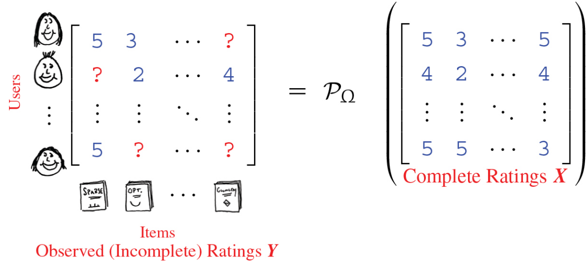
Recovering a low-rank matrix X_o:

$$\underbrace{y}_{\textit{observations}} = \mathcal{A} \left[\underbrace{X_o}_{\textit{unknown}} \right] \tag{22}$$

where $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map.

Examples of Low-rank Modeling

Recommendation Ratings:



$$\begin{bmatrix} 5 & 3 & \cdots & ? \\ ? & 2 & \cdots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & ? & \cdots & ? \end{bmatrix} = \mathcal{P}_\Omega \left(\begin{bmatrix} 5 & 3 & \cdots & 5 \\ 4 & 2 & \cdots & 4 \\ \vdots & \vdots & \ddots & \vdots \\ 5 & 5 & \cdots & 3 \end{bmatrix} \right)$$

We have:

$$\underbrace{Y}_{Observed \quad ratings} = \mathcal{P}_\Omega \left[\underbrace{X}_{Complete \quad ratings} \right] \tag{23}$$

where $\Omega := \{(i,j)|\text{user i has rated product j}\}$.

Examples of Low-rank Modeling

Many other examples::

- Multiple images of a Lambertian object
- Euclidean Distance Matrix Embedding
- Latent Semantic Indexing
- ...

Outline

Best Low-Rank Matrix Approximation

Theorem : Best Low-rank Approximation

Let $Y \in \mathbb{R}^{n_1 \times n_2}$, and consider the following optimization problem

$$\min_X \|Y - X\| \quad \text{such that } \operatorname{rank}(X) \leqslant r.$$

For any unitarily invariant (matrix) norm $\|.\|$, the optimal solution $\hat{X}$ has the form $\hat{X} = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, where $Y = \sum_{i=1}^{\min(n_1,n_2)} \sigma_i u_i v_i^T$ is the singular value decomposition of Y.

The same solution (truncating the SVD) applies to minimizing the rank of the unknown matrix X, subject to a data fidelity constraint:

$$\min_X \operatorname{rank}(X) \quad \text{such that } \|Y - X\| \leqslant \epsilon$$

Outline

General Rank Minimization

Problem: recover a low-rank matrix X from linear measurements:

$$\min_X \operatorname{rank}(X) \quad \text{s.t. } \mathcal{A}[X] = y$$

where $y \in \mathbb{R}^m$ is an observation and $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map:

$$\mathcal{A}[X] = (\langle A_1, X \rangle, ..., \langle A_m, X \rangle)^T, \quad A_i \in \mathbb{R}^{n_1 \times n_2}, \quad \langle P, Q \rangle = \operatorname{trace}(Q^T P)$$

[3]

[3] set of matrices A_i define our "measurements" y, through their inner (matrix) inner products with the unknown matrix X

General Rank Minimization

Problem: recover a low-rank matrix X from linear measurements:

$$\min_X \operatorname{rank}(X) \quad \text{s.t. } \mathcal{A}[X] = y$$

where $y \in \mathbb{R}^m$ is an observation and $\mathcal{A} \in \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^m$ is a linear map:

$$\mathcal{A}[X] = (\langle A_1, X\rangle, ..., \langle A_m, X\rangle)^T, \quad A_i \in \mathbb{R}^{n_1 \times n_2}, \quad \langle P, Q\rangle = \operatorname{trace}(Q^T P)$$

[3] Since $\operatorname{rank}(X) = \|\sigma(X)\|_0$, the problem is equivalent to the (NP-hard) ℓ^0 minimization:

$$\min_X \|\sigma(X)\|_0 \quad \text{s.t. } \mathcal{A}[X] = y$$

---

[3] set of matrices A_i define our "measurements" y, through their inner (matrix) inner products with the unknown matrix X

Convex Relaxation of Rank Minimization

Replace the rank, which is the ℓ^0 norm $\sigma(X)$ with the ℓ^1 norm of $\sigma(X)$:

$$\text{Nuclear norm: } \|X\|_* := \|\sigma(X)\|_1 = \textstyle\sum_i \sigma_i(X)$$

This is also known as the trace norm (for symmetric positive semidefinite matrices), the *Schatten 1-norm*, or the *Ky-Fan k-norm*.

Nuclear norm minimization problem:

$$\min_X \|X\|_* \qquad \text{subject to } \mathcal{A}[X] = y. \tag{24}$$

Nuclear Norm – Convex Envelope of Rank

Why $\|X\|_*$ is a norm (hence convex)?

Theorem

For $M \in \mathbb{R}^{n_1 \times n_2}$, let $\|M\|_* = \sum_{i=1}^{\min(n_1,n_2)} \sigma_i(M)$. Then $\|.\|_*$ is a norm.
Moreover, the nuclear norm and the the spectral norm are dual norms:

$$\|M\|_* = \sup_{\|N\|_2 \leqslant 1} \langle M, N \rangle, \quad \text{and } \|M\|_2 = \sup_{\|N\|_* \leqslant 1} \langle M, N \rangle \tag{25}$$

Why $\|X\|_*$ is tight to approximate rank(X)?

Theorem (Fazel 2002)

$\|M\|_*$ is the convex envelope of rank(M) over

$$\mathcal{B}_{op} := \{M | \|M\|_2 \leqslant 1\} \tag{26}$$

Nuclear Norm – Convex Envelope of Rank

Some remarks regarding the last theorem (FYI only)

- The theorem only applies to those M inside the unit ball. It tells nothing if M is outside the unit ball. In fact, if M is outside the unit ball, the output of the convex envelope will be at ∞.
- If we have $\mathcal{B}' := \{M \in \mathbb{R}^{n_1 \times n_2} \|M\|_2 \leqslant \alpha\}$, we can do a scaling to reuse the theorem : in this case the convex envelope of $\operatorname{rank}(M)$ on $\mathcal{B}'$ will be $1/\alpha \|M\|_*$ for $\alpha > 0$.
- The tool to prove the theorem is basically the convex conjugate, more particularly show that the *biconjugate* of the rank, that is $(\operatorname{rank}(X))^{**}$ is the nuclear norm of X.

Nuclear Norm – Variational Forms

How to compute besides SVD? (FYI only)

$\|X\|_*$ is equivalent to the following variational forms:

1. $\|X\|_* = \min_{U,V} \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2)$ s.t. $X = UV^T$.
2. $\|X\|_* = \min_{U,V} \|U\|_F \|V\|_F$ s.t. $X = UV^T$.
3. $\|X\|_* = \min_{U,V} \sum_k \|u_k\|_2 \|v_k\|_2$ s.t. $X = UV^T = \sum_k u_k v_k^T$.

These are useful in parameterizing low-rank matrices and finding them numerically, say via optimization.

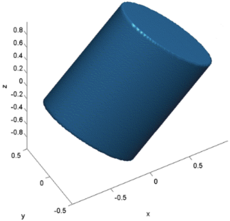
Success of Nuclear Norm – Geometric Intuition

Nuclear norm ball: consider the set of 2×2 symmetric matrices, parameterized as $M = \begin{pmatrix} x & y \\ y & z \end{pmatrix} \in \mathbb{R}^{2\times 2}$.

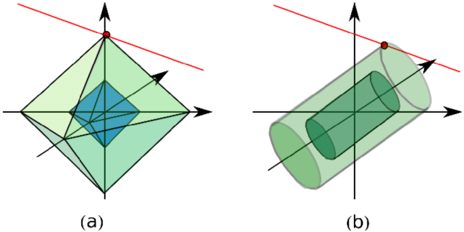
The nuclear norm (unit) ball

$$\mathcal{B}_* = \{M \,|\, \|M\|_* \leqslant 1\}$$

is a cylinder in $\mathbb{R}^3$!



The two circles at both ends of the cylinder correspond to matrices of rank 1.



Important theoretical guarantees skipped

You need to know that literature covers many theoretical aspects that we just list here for interested readers

1. the definition of the *Rank-Restricted Isometry Property*, defined for the operator $\mathcal{A}$ in our case.
2. *Rank Minimization Success*: conditions for the uniqueness of the optimal solution to the rank minimization problem, i.e. when the optimal solution of

$$\min_{X} \operatorname{rank}(X) \text{ s.t. } \mathcal{A}[X] = y$$

   denoted $\hat{X}$ is equal to X_o (the original X to recover).
3. *Nuclear Norm Minimization Success*: the conditions that guarantee that X_o is the unique optimal solution to the nuclear norm minimization problem

$$\min_{X} \|X\|_* \text{ s.t. } \mathcal{A}[X] = y$$

   .

Summary

Parallel developments for sparse vectors and low-rank matrices.

Sparse v.s. Low-rank	Sparse Vector	Low-rank Matrix
Low-dimensionality of	individual signal x	a set of signals X
Compressive sensing	$y = Ax$	$y = \mathcal{A}[X]$
Low-dim measure	ℓ^0 norm $\|x\|_0$	$\mathrm{rank}(X)$
Convex surrogate	ℓ^1 norm $\|x\|_1$	nuclear norm $\|X\|_*$

One can prove that nuclear norm minimization can recover w.h.p. a low-rank matrix X_o from $m = O(nr)$ random linear measurements $y = \mathcal{A}[X]$ (no proof :)).

Outline

Nuclear Norm Minimization

Problem (Matrix Completion)

Let $X_o \in \mathbb{R}^{n \times n}$ be a low-rank matrix. Suppose we are given $Y = \mathcal{P}_\Omega[X_o]$ where $\Omega \subseteq [n] \times [n]$. Fill in the missing entries of X_o.

Question: can we find X_o by solving the nuclear norm minimization:

$$\min_X \|X\|_* \text{ such that } \mathcal{P}_\Omega[X] = Y? \tag{27}$$

Simulations lead the way of investigation – need an algorithm...

Nuclear Norm Minimization

Remarks:

- Problem (27) is a special instance of the general nuclear norm minimization problem with observation operator $\mathcal{A} = \mathcal{P}_\Omega$.
- $\mathcal{P}$ is the projection operator onto the subset $\Omega \subseteq [n] \times [n]$ of the entries:

$$\mathcal{P}_\Omega[X](i,j) = \begin{cases} X_{ij} & \text{if } (i,j) \in \Omega, \\ 0 & \text{else.} \end{cases} \tag{28}$$

Algorithm via Augmented Lagrange Multiplier

Nuclear norm minimization for matrix completion:

$$\min_{X} \underbrace{\|X\|_*}_{f(x)} \text{ such that } \underbrace{\mathcal{P}_\Omega[X] = Y}_{g(x)=0} \tag{29}$$

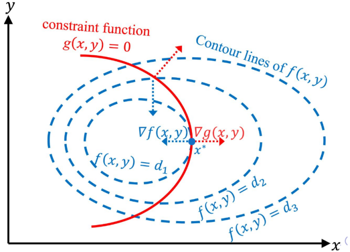
The Lagrangian method:

$$\mathcal{L}(X, \Lambda) = \|X\|_* + \langle \Lambda, Y - \mathcal{P}_\Omega[X] \rangle$$

Optimality conditions:

$$\frac{\partial \mathcal{L}}{\partial X} = 0, \frac{\partial \mathcal{L}}{\partial \Lambda} = 0 \tag{30}$$

However, it only holds at the point of the optimal solution $X^\star$..



Algorithm via Augmented Lagrange Multiplier

Instead of working with the Lagrangian function, we work with the so-called *Augmented* Lagrangian Function

$$\mathcal{L}_\rho(X, \Lambda) = \|X\|_* + \langle \Lambda, Y - \mathcal{P}_\Omega[X] \rangle + \frac{\rho}{2}\|Y - \mathcal{P}_\Omega[X]\|_F^2 \tag{31}$$

to derive more robustly convergence algorithm (see the final Optimization lecture)[4].

- The additional quadratic penalty term $\frac{\rho}{2}\|Y - \mathcal{P}_\Omega[X]\|_F^2$ encourages satisfaction of the constraint.
- The augmented Lagrangian method seeks for a saddle point of $\mathcal{L}_\rho$ by alternating between minimizing w.r.t. the "primal variables" X and taking one step of gradient ascent to increase $\mathcal{L}_\rho$ w.r.t. dual variables Λ:

$$\begin{aligned} \textbf{Primal}: &\quad X_{k+1} \in \operatorname{argmin}_X \mathcal{L}_\rho(X, \Lambda_k) \\ \textbf{Dual}: &\quad \Lambda_{k+1} := \Lambda_k + \rho \mathcal{P}_\Omega[Y - X_{k+1}] \end{aligned} \tag{32}$$

  Here $\mathcal{P}_\Omega[Y - X_{k+1}] = \nabla_\Lambda \mathcal{L}_\rho(X_{k+1}, \Lambda)$, and step size is set to ρ (choice is important !).

[4]intuitively, you may see the *augmented* Lagrangian as an attempt to regularize the landscape around the optimal solution $X^\star$

Algorithm: Proximal Gradient Descent

How to minimize the augmented Lagrangian $\mathcal{L}_\rho$:

$$\min_X F(X) := \underbrace{\|X\|_*}_{g(X) \text{ non-smooth convex}} + \underbrace{\langle \Lambda, Y - \mathcal{P}_\Omega[X] \rangle + \frac{\rho}{2}\|Y - \mathcal{P}_\Omega[X]\|_F^2}_{f(X) \text{ smooth, convex, } \rho\text{-Lipschitz}} \tag{33}$$

At each iterate X_k, construct a local (quadratic) upper bound for F:

$$\bar{F}(X, X_k) := g(X) + f(X_k) + \langle \nabla f(X_k), X - X_k \rangle + \frac{\rho}{2}\|X - X_k\|_F^2 \tag{34}$$

where $\nabla f(X) = -\mathcal{P}_\Omega[\Lambda] + \rho \mathcal{P}_\Omega[X - Y]$[5]

Proximal gradient descent: the next iterate X_{k+1} is computed as:

$$\begin{aligned} X_{k+1} &:= \text{argmin}_X \{\bar{F}(X, X_k)\} \\ &= \text{argmin}_X \{g(x) + \frac{\rho}{2}\|X - (\underbrace{X_k - \frac{1}{\rho}\nabla f(X_k)}_{M})\|_F^2\} \end{aligned} \tag{35}$$

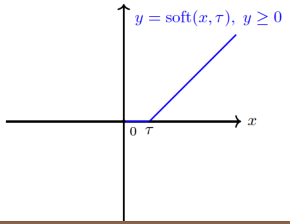
[5] you can show that the gradient is Lipschitz continuous with constant ρ

Algorithm: Proximal Operator for Nuclear Norm

For a matrix M with SVD $M = U\Sigma V^T$, its Singular Value Thresholding operator is:

$$\mathrm{SVT}_\tau[M] = US_\tau[\Sigma]V^T \tag{36}$$

where $S_\tau[X] = \mathrm{sign}(X) \odot (|X| - \tau)_+$ is the entry-wise soft thresholding operator.



Theorem

The unique solution $X^\star$ to the problem:

$$\min_X \{\|X\|_* + \frac{\rho}{2}\|X - M\|_F^2\} \tag{37}$$

is given by

$$X^\star = \mathrm{SVT}_{\frac{1}{\rho}}[M] = U[\Sigma - \frac{I}{\rho}]_+ V^T, \quad [.]_+ = \max(.,0). \tag{38}$$

What does SVT ?

1. Compute $M := X_k - \frac{1}{\rho}\nabla f(X_k)$
2. Perform SVD on M and get $U\Sigma V^T$
3. Subtract all the diagonal value of Σ by $\frac{1}{\rho}$, denoted $\Sigma - \frac{I}{\rho}$
4. Replace negative value in $\Sigma - \frac{I}{\rho}$ by zero, denoted $[\Sigma - \frac{I}{\rho}]_+$
5. Compute $U[\Sigma - \frac{I}{\rho}]_+ V^T$

Algorithm via Augmented Lagrange Multiplier

Outer Loop: Matrix Completion by ALM

input : $X_0 = \Lambda_0 = 0$, $\rho > 0$, and Y.

while *not converged* **do**

  compute $X_{k+1} \in \operatorname{argmin}_X \mathcal{L}_\rho(X, \Lambda_k)$ (say by PG)

  compute $\Lambda_{k+1} := \Lambda_k + \rho(Y - \mathcal{P}_\Omega[X_{k+1}])$

end

Inner Loop: Proximal Gradient

input : X_0 starts with the X_k, and $\Lambda := \Lambda_k$ from the outer loop, and Y.

while *not converged* **do**

  compute $\nabla f(X_l) = -\mathcal{P}_\Omega[\Lambda] + \rho \mathcal{P}_\Omega[X_l - Y]$

  compute $M := X_l - \frac{1}{\rho} \nabla f(X_l)$

  compute $M = U \Sigma V^T$

  compute $X_{l+1} := \mathrm{SVT}_{\frac{1}{\rho}}[M] = U[\Sigma - \frac{I}{\rho}]_+ V^T$

end

When Nuclear Norm Minimization Succeeds ?

To understand when the convex optimization Problem (27) and when the above algorithm correctly recover a matrix $X = X_o$ from a part of its entries:

- we vary the rank r of the matrix X_o as a fraction of the dimension n
- and a fraction $p \in (0, 1)$ of (randomly chosen) chosen entries.
  In other words, p is the probability that an entry is given.

When Nuclear Norm Minimization Succeeds?

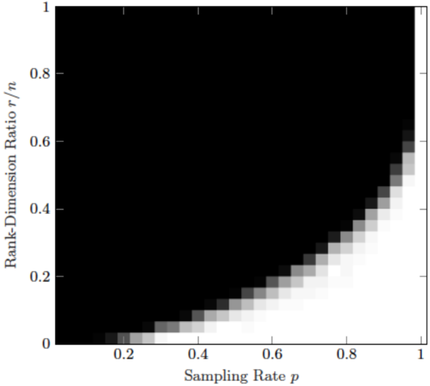


Figure: Matrix completion for varying rank and sampling rate. Fraction of correct recoveries across 50 trials, as a function of the rank-dimension ratio r/n (vertical axis) and fraction p of observed entries (horizontal axis). Here, $n = 60$. In all cases, $X_o = AB^T$ is a product of two independent $n \times r$ i.i.d. $\mathcal{N}(0, 1/n)$ matrices. Trials are considered successful if $\frac{\|\hat{X} - X_o\|_F}{\|X_o\|_F} \leqslant 10^{-3}$.

When Nuclear Norm Minimization Succeeds?

Few observations from the simulations:

1. the convex optimization Problem (27) and the Algorithm indeed succeed under a surprisingly wide range of conditions, as long as the rank of the matrix is relatively low and a (sufficient) fraction of the entries are observed.
2. the success and failure exhibit a sharp phase transition phenomenon.

▸ Let us check together these results !

When Nuclear Norm Minimization Succeeds?

When it fails?

1. if X_o is itself sparse
2. if Ω is chosen adversarially (e.g., an entire row or column of X_o).

When Nuclear Norm Minimization Succeeds?

Notice for any rank-r orthogonal matrix U:

$$\sum_i \|e_i^T U\|_2^2 = \|U\|_F^2 = r \rightarrow \max_i \|e_i^T U\|_2^2 \geqslant r/n \tag{39}$$

Definition

We say that $X_o = U\Sigma V^T$ is ν-incoherent if the following hold:

$$\begin{aligned} \forall i \in [1,n], \quad & \|e_i^T U\|_2^2 \leqslant \nu r/n, \\ \forall j \in [1,n], \quad & \|e_j^T V\|_2^2 \leqslant \nu r/n. \end{aligned} \tag{40}$$

These two conditions control the "spikiness" of the singular vectors of X_o. If ν is small, the singular are spread around.

When Nuclear Norm Minimization Succeeds?

Bernoulli Ber(p) **sampling model**: each entry (i, j) belongs to the observed set Ω independently with probability $p \in [0, 1]$. Hence, the expected number of observed entries is:

$$m = \mathbb{E}[|\Omega|] = pn^2 \tag{41}$$

Theorem:Matrix Completion via Nuclear Norm Minimization

Let $X_o \in \mathbb{R}^{n \times n}$ be a rank-r matrix with incoherence parameter ν. Suppose that we observe $Y = \mathcal{P}_\Omega[X_o]$, with Ω sampled according to the Bernoulli model with probability

$$p \geqslant C_1 \frac{\nu r \log^2(n)}{n}$$

Then, with probability at least $1 - C_2 n^{-c_3}$, X_o is the unique optimal solution to

$$\min_X \|X\|_* \text{ subject to } Y = \mathcal{P}_\Omega[X]$$

In brief: we need $m = pn^2 = O(nr \log^2(n))$ randomly sampled entries: $Y = \mathcal{P}_\Omega[X]$.

Summary

Summary

We have seen :

- Geometric intuition of why ℓ^1 Minimization succeeds to recover sparse signal x_o.
- Correctness of ℓ^1 Minimization with conditions on $\mu(A)$ (Mutual Coherence of the matrix A).
- Construct incoherent matrices (with small $\mu(A)$), and limitations of Incoherence: recovering k-sparse solution requires $m = O(k^2)$ measurements, while experimental results seem to indicate less.
- General Rank Minimization: recover a low-rank matrix X from linear measurements, and its convex relaxation, the *Nuclear Norm Minimization*.
- A SOTA Algorithm for solving the Nuclear Norm Minimization via *Augmented Lagrangian function*.
- Some comments and results on the success of the nuclear norm min.

Preparation for the lab

- Review the lecture :).
- Implement the Algorithm in Python for Matrix Completion.

Decomposing Low-Rank and Sparse Matrices

Principal Component Pursuit: Algorithms

Under construction

Goodbye, So Soon

THANKS FOR THE ATTENTION

- v.leplat@innopolis.ru
- sites.google.com/view/valentinleplat/