On the polynomial complexity of exact recovery

Stephen Vavasis[1]

[1]Department of Combinatorics & Optimization
University of Waterloo

Parts of this talk represent joint work with X. V. Doan of Warwick and K.-C. Toh of N. U. Singapore

2012-Nov-20 / Workshop on Conic Programming

Machine learning

- Until about 1990, machine learning was dominated by logic and rule-based reasoning.
- E.g., for text processing, make rules for how parts of speech interact.
- Starting around 1990, paradigm shift in ML to data mining and statistics based on large training sets.
- Computational problem: finding patterns in large data sets.

Machine learning (cont'd)

- Often the computational problems of interest, such as finding a dense cluster of nodes in a large sparse graph, are NP-hard.
- Yet the problems are routinely solved satisfactorily in practice using heuristics.
- Suggests that real data has hidden structure that makes finding patterns easier.

Generative models

- How to model this hidden structure? One popular approach: generative model.
- Assume that the data is produced by a process involving deterministic (adversary-based) choices and random numbers.
- Try to prove that a particular algorithm can solve problems produced by the model in polynomial time with high probability

Recent successes with convex optimization

Convex programming exactly solves many NP-hard data mining problems in polynomial time when the instance comes from a generative model:

- Compressive sensing (Donoho; Candès, Romberg & Tao)
- Rank minimization (Recht, Fazel, Parrilo)
- Matrix completion (Candès & Recht; Candes & Tao)
- Rank-sparsity decomposition (Chandrasekaran et al.; Candès et al.)
- Clique & clustering (Ames & V.)
- Nonnegative matrix factorization (Doan, Toh & V)

Is it really polynomial time?

- Except for LP, exact solution to SDP not attainable. Even for LP, complexity issues must be resolved.
- Not obvious that an exact solution to the original problem is obtained from an approximate solution to the convex relaxation. And how approximate?
- Thus, it is fair to ask whether the above results are truly exactly solving the original problem in polynomial time. (Y. Ye)

Compressive sensing

Compressive sensing LP, $\min \|\mathbf{x}\|_1$ s.t. $A\mathbf{x} = \mathbf{b}$, involves coefficient matrices A that are typically Bernoulli, Gaussian or random Fourier.

- Bernoulli: number of bits L to write the input is $\mathrm{poly}(m, n)$. Thus, ellipsoid or interior point always polynomial time for these cases.
- Fourier: also $\mathrm{poly}(m, n)$ (Adler & Beling, 1991)
- Gaussian: Tuncel, Todd & Ye (2001) show that V.-Ye interior point method (real-number arithmetic) solves LP exactly in $\mathrm{poly}(m, n)$ time with probability very close to 1.

Other choices of A apparently need case-by-case analysis.

SDP case

- Focus on a particular problem and algorithm: work by Doan, Toh & V. on nonnegative matrix factorization.
- Attempt to broaden the idea.

Finding a feature in a text dataset

Suppose one is given a *text corpus*, i.e., a collection of n text documents, and one seeks a topic in the dataset, that is, a subset of related documents. One approach:

- Form the term-document matrix, that is, the $m \times n$ matrix in which ith row corresponds to the ith term, jth column to jth document, and $A(i,j)$ is the number of occurrences of term i in document j.
- Find a large approximately rank-one submatrix $A(I,J)$ of A (i.e., $A(I,J) \approx \mathbf{w}\mathbf{h}^T$).

Finding a feature in an image dataset

Given an image dataset in which all the n contain exactly $m_1 \times m_2 \equiv m$ pixels, find a visual feature, that is, a particular pattern that recurs in the same subset of pixels in a subset of images.

- Form an $m \times n$ matrix A in which $A(i,j)$ stands for the intensity of pixel i in image j.
- Find a large approximately rank-one submatrix (LAROS) $A(I,J)$ of A.

LAROS and NMF

- Assume A is nonnegative.
- The above process can be repeated iteratively:

  For $i = 1 : k$

  Find $I_i, J_i, \bar{\mathbf{w}}_i, \bar{\mathbf{h}}_i$ s.t. $A(I_i, J_i) \approx \bar{\mathbf{w}}_i \bar{\mathbf{h}}_i^T$.

  Pad $(\bar{\mathbf{w}}_i, \bar{\mathbf{h}}_i)$ with zeros to obtain $(\mathbf{w}_i, \mathbf{h}_i)$.

  $A = \max(A - \mathbf{w}_i \mathbf{h}_i^T, 0)$.
- Upon completion,
  $A \approx \mathbf{w}_1 \mathbf{h}_1^T + \cdots + \mathbf{w}_k \mathbf{h}_k^T \equiv WH^T$.

Greedy NMF algorithm

- OK to assume that $\mathbf{w}_i \geq \mathbf{0}$, $\mathbf{h}_i \geq \mathbf{0}$ (Perron-Frobenius).
- Given a nonnegative matrix A, a factorization $A \approx WH^T$ is called *nonnegative matrix factorization* (NMF) if W, H both nonnegative.
- The algorithm on the previous transparency is a greedy NMF algorithm (Asgarian & Greiner, Bergmann et al., Biggs et al., Gillis & Glineur).

LAROS and SVD

- Best overall rank-one approximation to A comes from SVD (Eckart-Young theorem).

$$A = \begin{pmatrix} 0.8 & 0.9 & 0.0 & 0.0 \\ 0.8 & 1.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.9 \\ 0.0 & 0.0 & 1.1 & 0.8 \end{pmatrix}.$$

- The dominant left singular vector is $\approx [1; 1; 0; 0]$; SVD has identified $A(1:2, 1:2)$.
- But with a little noise, dominant left singular vector $\approx [1; 1; 1; 1]$; SVD fails to identify LAROS.

LAROS and SVD

- Best overall rank-one approximation to A comes from SVD (Eckart-Young theorem).

$$A = \begin{pmatrix} 0.8 & 0.9 & \mathbf{0.1} & \mathbf{0.2} \\ 0.8 & 1.1 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.8 & 0.9 \\ 0.0 & 0.0 & 1.1 & 0.8 \end{pmatrix}.$$

- The dominant left singular vector is $\approx [1; 1; 0; 0]$; SVD has identified $A(1:2, 1:2)$.
- But with a little noise, dominant left singular vector $\approx [1; 1; 1; 1]$; SVD fails to identify LAROS.

SVD as optimization

- The solution to this problem is to modify the SVD to promote sparsity.
- Can write SVD as an optimization problem (Eckart-Young) and add another term, i.e.,

$$\min_{\sigma,\mathbf{u},\mathbf{v}} \|A - \sigma \mathbf{u}\mathbf{v}^T\| + \text{densityPenalty}(\mathbf{u},\mathbf{v})$$

- Unfortunately, Eckart-Young optimization problem is not convex.

SVD as convex optimization

- Let $\|\cdot\|_*$ denote the *nuclear norm*, that is, $\|X\|_* = \sigma_1(X) + \cdots + \sigma_n(X)$.
- Theorem: The nuclear norm is dual to the 2-norm, i.e., $\|X\|_* = \max\{Z \bullet X : \|Z\|_2 \leq 1\}$.
- Given A, the solution to the convex optimization problem $\min\{\|X\|_* : A \bullet X \geq 1\}$ is $X = \mathbf{u}_1\mathbf{v}_1^T/\sigma_1$, where $(\sigma_1, \mathbf{u}_1, \mathbf{v}_1)$ is the dominant singular triple of A.

Obtaining a sparse solution

- In order to enforce sparsity, could add a (nonconvex) penalty term: $\min \|X\|_* + \pi(|I| \cdot |J|)$ s.t. $A \bullet X \geq 1; (i,j) \notin I \times J \Rightarrow X(i,j) = 0$. where $\pi(\cdot)$ is an increasing penalty function.
- The optimal X will have necessarily have the form $X = \bar{\mathbf{u}}_1 \bar{\mathbf{v}}_1^T / \bar{\sigma}_1$, where $(\bar{\sigma}_1, \bar{\mathbf{u}}_1, \bar{\mathbf{v}}_1)$ is the dominant singular triple of $A(I,J)$ for some (I,J) padded with zeros.
- This problem is NP-hard.

Convex relaxation of sparsity

- A common technique in the literature to promote sparsity is adding an ℓ_1 penalty term.
- Applying this to the preceding nonconvex problem yields

$$\begin{array}{ll} \min & \|X\|_* + \theta\|X\|_1 \\ \text{s.t.} & A \bullet X \geq 1. \end{array}$$

- Note: $\|X\|_1$ means $\|\mathrm{vec}(X)\|_1$;
- Above problem is convex. (Indeed, it is semidefinite programming.)
- Nuclear-plus-1-norm has also appeared in rank-sparsity decomposition work.

Recoverability

- Suppose $A \geq 0$ has the form $A = \mathbf{u}\mathbf{v}^T + R$ where $\mathbf{u}$,$\mathbf{v}$ are sparse and R is random noise. Can we recover $(\mathbf{u}, \mathbf{v})$ from A?
- No, but maybe we can recover $\mathrm{supp}(\mathbf{u})$ and $\mathrm{supp}(\mathbf{v})$ (positions of nonzero entries).
- Assume that R is i.i.d. random. Assume $\mathbf{u}, \mathbf{v}$ are deterministic and positive.

Recoverability

- Suppose $A \geq 0$ has the form $A = \mathbf{u}\mathbf{v}^T + R$ where $\mathbf{u}$,$\mathbf{v}$ are sparse and R is random noise. Can we recover $(\mathbf{u}, \mathbf{v})$ from A?
- No, but maybe we can recover $\mathrm{supp}(\mathbf{u})$ and $\mathrm{supp}(\mathbf{v})$ (positions of nonzero entries).
- Assume that R is i.i.d. random. Assume $\mathbf{u}, \mathbf{v}$ are deterministic and positive.

Main theorem on recoverability

- Say $A \in \mathbf{R}^{M \times N}$; $|\operatorname{supp}(\mathbf{u})| = m$; $|\operatorname{supp}(\mathbf{v})| = n$.
- Assume entries of R are i.i.d. subgaussian about their mean μ.
- Assume the mean of R is bounded in terms of the divergence of $\mathbf{u}, \mathbf{v}$ from $\mathbf{e}$.
- Assume θ chosen in a certain range.
- Then convex relaxation recovers $\operatorname{supp}(\mathbf{u}), \operatorname{supp}(\mathbf{v})$ with prob. exponentially close to 1 provided $m \geq \Omega(\sqrt{M})$ and $n \geq \Omega(\sqrt{N})$.

Proof steps

- To simplify notation, assume support of $\mathbf{u}, \mathbf{v}$ are their leading indices.
- Hypothesize existence of optimal solution of the form

$$X = \begin{pmatrix} \sigma_1 \bar{\mathbf{u}} \bar{\mathbf{v}}^T & 0 \\ 0 & 0 \end{pmatrix},$$

  $\|\bar{\mathbf{u}}\| = \|\bar{\mathbf{v}}\| = 1.$
- KKT condition is $\lambda A = Y + \theta Z$ for some $Y \in \partial \|X\|_*$, $Z \in \partial \|X\|_1$, $\lambda \geq 0$.
- KKT condition sufficient for global optimality in convex optimization.

Proof steps (cont'd)

- $\lambda A = Y + \theta Z$ for some $Y \in \partial\|X\|_*$, $Z \in \partial\|X\|_1$, $\lambda \geq 0$.
- Specializing to preceding X this means: dominant singular triple of Y is $(1, [\bar{\mathbf{u}}; \mathbf{0}], [\bar{\mathbf{v}}; \mathbf{0}])$; $\|Z\|_\infty = 1$ and $Z_{11} = \mathtt{ones}(\mathtt{m}, \mathtt{n})$.
- Implies that λ must be chosen so that $\|\lambda A_{11} - \theta \cdot \mathtt{ones}(\mathtt{m}, \mathtt{n})\| = 1$.
- This is an algebraic equation for λ; can get good estimates for λ because there is a good upper bound known for the norm of a mean-zero random matrix.

Proof steps (cont'd)

- Once λ is known, $\bar{\mathbf{u}}, \bar{\mathbf{v}}$ are dominant singular vectors of $\lambda A_{11} - \theta \cdot \mathtt{ones}(\mathtt{m}, \mathtt{n})$.
- With these choices for $\lambda, \bar{\mathbf{u}}, \bar{\mathbf{v}}$, must next fill in the rest of Y and Z so that $\|Y\| \leq 1$ and $\|Z\|_\infty \leq 1$.
- The requirement $\|Y\| \leq 1$ couples the four blocks together, so replace it with the restriction that $\|Y_{ij}\| \leq 1/2$ for $i, j = 1, 2$.

Proof steps (cont'd)

- KKT multipliers Y_{22} and Z_{22} constructed by taking the mean of λA into Z_{22} (i.e., make it a multiple of the all-1's matrix) and deviations from average in Y_{22}. Uses the fact that $\|R\|$ is (unexpectedly?) small when R is a random mean-0 matrix.
- Construction of KKT multipliers Y_{12}, Z_{12} are more complicated because condition on dominant singular triple of Y imposes linear constraint $\bar{\mathbf{u}}^T Y_{12}=0$.
- Need estimates of $\bar{\mathbf{u}}, \bar{\mathbf{v}}$; use Wedin's sine theorem (SVD perturbation theorem).

Recovery of $\mathrm{supp}(\mathbf{u}), \mathrm{supp}(\mathbf{v})$

- The proof of the theorem shows that, under the assumptions and with high probability, $\mathrm{rank}(X) = 1$, i.e., $X = \hat{\mathbf{u}}\hat{\mathbf{v}}^T$ where $\hat{\mathbf{u}}$ is the extension of $\bar{\mathbf{u}}$ with zeros and similarly for $\hat{\mathbf{v}}$.
- Furthermore, $\mathrm{supp}(\mathbf{u}) = \mathrm{supp}(\hat{\mathbf{u}})$ and $\mathrm{supp}(\mathbf{v}) = \mathrm{supp}(\hat{\mathbf{v}})$.

Convex solver

- Recall our relaxation

$$\begin{array}{ll} \min & \|X\|_* + \theta \|X\|_1 \\ \text{s.t.} & A \bullet X \geq 1. \end{array}$$

  is convex and indeed SDP-expressible.
- Interior point SDP solvers (Sedumi, SDPT3) require $O(p^3)$ flops per iteration, where $p = MN$ (number of unknowns).
- Interior point methods give accuracy ϵ after $\text{poly}(\log(1/\epsilon))$ iterations.
- Too inefficient for large problems.

Subgradient descent

- We use a subgradient descent method.
- On each step, approximately minimize *proximal point mapping*. Proximal-point mapping for convex $\phi(\mathbf{x})$ defined to be solution to $\min_{\mathbf{x}} \phi(\mathbf{x}) + \frac{\lambda}{2}\|\mathbf{x} - \mathbf{c}\|^2$ (2-norm for vectors, F-norm for matrices).

Proximal point mapping and new termination test

- We do not know how to efficiently minimize the proximal-point mapping for our objective function $\phi(X) = \|X\|_* + \theta\|X\|_1 + \frac{\lambda}{2}\|X - C\|_F^2$.
- Therefore, rewrite relaxation as

$$\begin{array}{rl} \min & \|X_1\|_* + \theta\|X_2\|_1 \\ \text{s.t.} & A \bullet X_1 \geq 1, \\ & X_1 = X_2 \end{array}$$

- This allows us to compute the proximal point mapping separately for $\|\cdot\|_*$ and $\|\cdot\|_1$.

Proximal point mapping for nuclear norm

- Proximal-point mapping for nuclear norm: given C, minimizer of $\|X\|_* + \frac{\lambda}{2}\|X - C\|_F^2$ is

$$U \begin{pmatrix} (\sigma_1 - 1/\lambda)^+ & & \\ & \ddots & \\ & & (\sigma_n - 1/\lambda)^+ \end{pmatrix} V^T,$$

where $C = U\Sigma V^T$.

- Proximal point algorithm requires $\mathrm{poly}(1/\epsilon)$ iterations

Computational experiments

- Two black/white image datasets used in experiments.
- In both cases, LAROS run repeatedly in order to extract several features (find approximate NMF).
- Termination test: either as on previous transparency, or achievable accuracy achieved.
- Choice of θ: heuristic used.

Frey face data

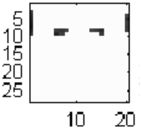
Frey face dataset consists of 1965 grayscale mugshots of a person's face in different poses.

Applying the method to Frey dataset

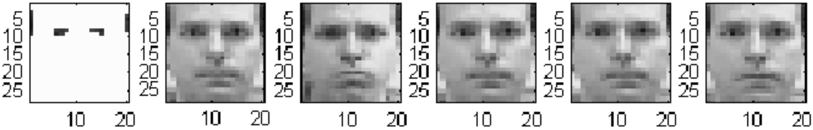
- Can form a 560×1965 matrix, one mugshot per column and look for a large rank-one submatrix.
- Feature corresponds to subset of images in database with common visual feature in the same groups of pixels.
- Can find multiple features by iteratively solving LAROS and subtracting off previous features.

Results

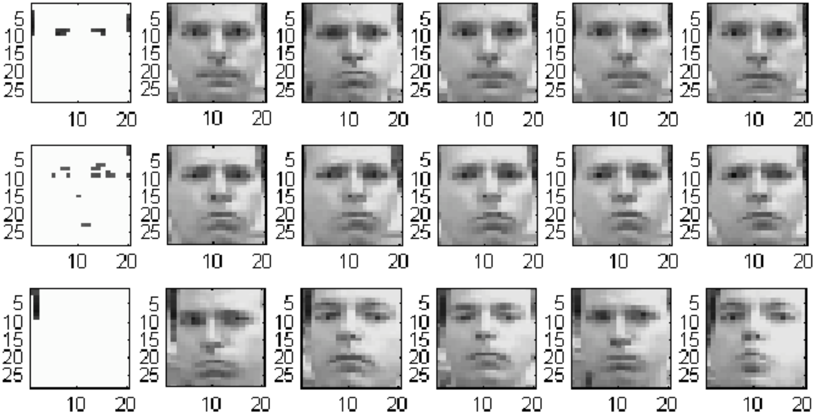
Results



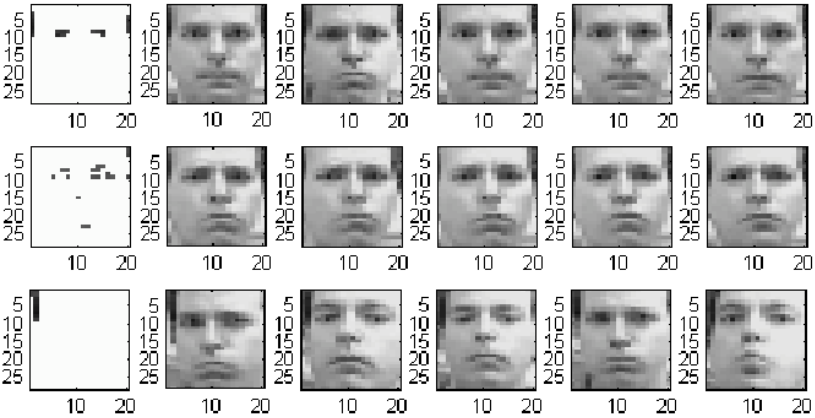
Results



Results



Results



This SDP has $> 10^6$ variables.

Termination test

- Since only nonzero pattern of optimal X^* is useful, would like to terminate as soon as nonzero pattern is determined.
- Test should also confirm that $\mathrm{rank}(X^*) = 1$. (If this equation fails, then exact recovery not possible.)
- Would like a test that, when satisfied, certifies that correct answer has been found.

Nonlinear equations

- Given approximate solution $\tilde{X}$, find approximate dominant singular triple $(\tilde{\sigma}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})$ and Lagrange multiplier $\tilde{\lambda}$.
- Consider system of equations:

$$\begin{aligned} (\lambda A_{11} - \theta Z_{11})\mathbf{v} &= \mathbf{u}, \\ (\lambda A_{11} - \theta Z_{11})^T\mathbf{u} &= \mathbf{v}, \\ \mathbf{u}^T A_{11}\mathbf{v} &= 1. \end{aligned}$$

  where Z_{11} is all 1's.
- First two express the fact that $(1, \mathbf{u}_1, \mathbf{v}_1)$ form a singular-vector triple; last is nonstandard normalization.

Kantorovich theorem

- Can apply Kantorovich theorem to certify that the system has an exact solution distance ϵ from $(\tilde{\lambda}, \tau\tilde{\mathbf{u}}, \tau\tilde{\mathbf{v}})$.
- KKT conditions for a rank-one sparse solution include above equations and also inequalities.
- Use simple least squares to guess remaining multipliers.
- Check whether the inequalities hold for all points within a ball of radius ϵ around $(\tilde{\lambda}, \tilde{\mathbf{u}}, \tilde{\mathbf{v}})$.
- If so, a rank-one solution with correct sparsity pattern is guaranteed.

Complexity implication

- Can carry out *a priori* analysis to determine when termination test will be satisfied for data from generative model.
- Three requirements in Kantorovich theorem for certifying existence of nearby exact solution: $(\lambda, \mathbf{u}, \mathbf{v})$:
  - $\|P(\lambda, \mathbf{u}, \mathbf{v})\|$ should be small;
  - $\|\nabla P(\lambda, \mathbf{u}, \mathbf{v})^{-1}\|$ should be modest; and
  - $\|\nabla^2 P\|$ should be modest;

  where

$$P(\lambda, \mathbf{u}, \mathbf{v}) = \begin{pmatrix} (\lambda A_{11} - \theta Z_{11})\mathbf{v} - \mathbf{u} \\ (\lambda A_{11} - \theta Z_{11})^T \mathbf{u} - \mathbf{v} \\ \mathbf{u}^T A_{11} \mathbf{v} - 1 \end{pmatrix}.$$

Analysis of first requirement

"$\|P(\lambda, \mathbf{u}, \mathbf{v})\|$ should be small":

- Third equation of $P(\lambda, \mathbf{u}, \mathbf{v}) = \mathbf{0}$ is exact after scaling.
- The first two express the condition that $(1, \mathbf{u}, \mathbf{v})$ form a SVD triple of $\lambda A_{11} - \theta Z_{11}$.
- Wedin sine theorem states that perturbing the matrix by a small amount also perturbs the singular vectors by a small amount, assuming strict separation of singular values.
- Doan and V. show that in the proposed generative model, the second singular value of $\lambda A_{11} - \theta Z_{11}$ is $\leq 1/2$.

Analysis of third requirement

"$\|\nabla^2 P\|$ should be modest":

- Observe that P is quadratic hence $\nabla^2 P$ is a constant: involves only A_{11}.
- The norm of A_{11} is bounded in the generative model.

Analysis of second requirement

"$\|\nabla P(\lambda, \mathbf{u}, \mathbf{v})^{-1}\|$ should be modest":

- $\nabla P(\lambda, \mathbf{u}, \mathbf{v})$ has 3×3 block structure and is symmetric.
- Inverse can be analyzed using block Gaussian elimination; eliminate $\mathbf{u}$ then λ.
- Only complication is inverse S^{-1} of Schur complement, $S = I - B^T B + \mathbf{g}\mathbf{g}^T$ where $B = \lambda A_{11} - \theta Z_{11}$ and $\mathbf{g} = (A_{11}^T \mathbf{u} + B^T A_{11} \mathbf{v}) / \|A_{11} \mathbf{v}\|$.

Analysis of $S = I - B^T B + \mathbf{g}\mathbf{g}^T$

- $\|B\| \approx 1$; other singular vals $\leq 1/2 + \epsilon$; so $I - B^T B$ has one eigenvalue close to 0 and the others strictly positive
- Thus, can argue that $S \geq \delta I$ provided $\mathbf{g}$ has a big component in the eigenvector of $B^T B$ whose eigenvalue is 1.
- At the solution, this eigenvector is $\bar{\mathbf{v}}$. Therefore $(\bar{\mathbf{v}}^T\mathbf{g} \approx \bar{\mathbf{v}}^T A_{11}^T \bar{\mathbf{u}} + \bar{\mathbf{v}}^T B^T A_{11}\bar{\mathbf{v}})/\|A_{11}\bar{\mathbf{v}}\| = 2\bar{\mathbf{v}}^T A_{11}^T \bar{\mathbf{u}}/\|A_{11}\bar{v}\|$.
- This quantity can be lower-bounded by positivity.

Summary of this analysis

- Analysis shows that Kantorovich requirements will be satisfied when solution is within $1/\text{poly}(m, n)$ of optimizer.
- This is polynomial-time even for first-order methods that have sublinear convergence.
- Analysis showing that convex relaxation exactly solves original problem also applies to Kantorovich test.

Other possible applications

- Consider e.g. the matrix completion problem: given partially specified matrix $M \in \mathbf{R}^{m \times n}$ such that M_{ij} known whenever $(i, j) \subset \Omega$ (Ω sparse subset of $\{1, \ldots, m\} \times \{1, \ldots, n\}$), find the lowest rank $X \in \mathbf{R}^{m \times n}$ such that $X_{ij} = M_{ij}$ for all $(i, j) \in \Omega$.
- Solved in polynomial time via convex relaxation $\min \|X\|_* \text{ s.t. } X_{ij} = M_{ij} \forall (i, j) \in \Omega$ (Candès & Recht; Candès & Tao) assuming (Ω, M_Ω) generated according to a certain model.
- Can $\mathrm{rank}(X)$ be determined in polynomial time from an approximate solution to the convex problem?

KKT condition

- KKT conditions for relaxation are $X|_\Omega = M|_\Omega$, $G \in \partial\|X\|_*$, $G|_{\bar{\Omega}} = 0$. Here $\bar{\Omega}$ denotes the complement of Ω.
- The condition $G \in \partial\|X\|_*$ means that $\|G\| = 1$ and that left and right singular subspaces associated with $\sigma_{\max} = 1$ contain the spans of X and X^T resp.

KKT condition in rank-one case

- In the case $\mathrm{rank}(X) = 1$, these conditions imply that the following equations hold:

$$\begin{aligned} G\mathbf{v} &= \mathbf{u} \\ G^T\mathbf{u} &= \mathbf{v} \\ \mathbf{u}\mathbf{v}^T|_\Omega &= M|_\Omega \end{aligned}$$

- Here, $\mathbf{u}$ and $\mathbf{v}$ are rescalings of the nonvanishing left and right singular vectors of X.
- This is a square nonlinear system: $m + n + |\Omega|$ equations and and equal number of variables (only the nonzero positions of G are variables).

Open questions

- For MCP: generalize Kantorovich equation to $\mathrm{rank}(X) \geq 2$; prove polynomial-time recovery of rank.
- General recipe for termination tests and certificates of exact recovery?
- For compressive sensing, do RIP/width/CS-1..3 assumptions also imply polynomial-time exact LP solution?
- Make the tests efficient?