Computational Intelligence Laboratory

Lecture 3

Matrix Approximation & Reconstruction

Thomas Hofmann

ETH Zurich – cil.inf.ethz.ch

8 March 2019

Section 1

Collaborative Filtering

Collaborative Filtering

- **Recommender systems**
  - analyze patterns of interest in items (products, movies, ...)
  - provide personalized recommendations for users
- **Collaborative Filtering**
  - exploit collective data from many users
  - generalize across users and – possibly – across items
- Applications:
  - Amazon, Netflix, Pandora, online advertising, etc.
  - special case of **algorithmic selection**

Netflix Data

- Input: user-item preferences stored in a matrix
  - **rows = users**, **columns = items**

18,000 movies

480,000 users

x	1	1	x	...	x
x	x	x	5	...	x
x	x	3	x	...	x
x	4	3	x	...	2
...	x	x	x	...	x
x	5	x	1	...	x
x	x	3	3	...	x
x	1	x	x	...	2

  - 1-5 star rating of movies. x denotes a missing value.
  - predict missing values = matrix completion

Matrix Completion

- How can we fill in missing values?
- **Statistical model** with $k \ll m \cdot n$ parameters
  - $m \times n$: dimensionality of rating matrix
  - introduces coupling between entries
  - infer missing entries from observed ones
- **Low Rank** decomposition
  - find best approximation with low rank r
  - entries in decomposition: $k \leq r \cdot (m + n)$

Section 2

Matrix Approximation via SVD

Frobenius Norm: revisted

- Definition: **Frobenis norm**

$$\|\mathbf{A}\|_F := \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} a_{ij}^2} = \|\mathrm{vec}(\mathbf{A})\|_2 = \sqrt{\mathrm{trace}(\mathbf{A}^\top\mathbf{A})}$$

- Frobenius norm only depends on singular values of $\mathbf{A}$

$$\|\mathbf{A}\|_F^2 = \sum_{i=1}^{k} \sigma_i^2, \quad k = \min\{m, n\}$$

  - follows from **cyclic property**: $\mathrm{trace}(\mathbf{XYZ}) = \mathrm{trace}(\mathbf{ZXY})$

$$\mathrm{trace}(\mathbf{A}^\top\mathbf{A}) = \mathrm{trace}(\mathbf{V}\mathbf{D}^\top\mathbf{U}^\top\mathbf{U}\mathbf{D}\mathbf{V}^\top) = \mathrm{trace}(\mathbf{V}^\top\mathbf{V}\mathbf{D}^\top\mathbf{D})$$
$$= \mathrm{trace}(\mathbf{D}^\top\mathbf{D}) = \mathrm{trace}(\mathrm{diag}(\sigma_1^2 \ldots, \sigma_k^2)) = \sum_{i=1}^{k} \sigma_i^2$$

Singular Values and Matrix Norms

- **Induced p-norms**

$$\|\mathbf{A}\|_p := \sup\{\|\mathbf{A}\mathbf{x}\|_p : \|\mathbf{x}\|_p = 1\}, \quad \|\mathbf{x}\|_p := \left(\sum_i |x_i|^p\right)^{1/p}$$

- Matrix 2-norm (**spectral norm**) = largest singular value

$$\|\mathbf{A}\|_2 = \sup\{\|\mathbf{A}\mathbf{x}\|_2 : \|\mathbf{x}\|_2 = 1\} = \sigma_1$$

  - assume $\|\mathbf{x}\|_2 = 1$, define $\mathbf{y} := \mathbf{V}^\top \mathbf{x}$, then $\|\mathbf{y}\|_2 = 1$ ($\mathbf{V}$ orthogonal)
  - define $\mathbf{z} := \mathbf{D}\mathbf{y}$, then $\|\mathbf{A}\mathbf{x}\|_2 = \|\mathbf{U}\mathbf{z}\|_2 = \|\mathbf{z}\|_2$ ($\mathbf{U}$ orthogonal)
  - hence: $\|\mathbf{A}\mathbf{x}\|_2^2 = \|\mathbf{D}\mathbf{y}\|_2^2 = \sum_{i=1}^k \sigma_i^2 y_i^2$
  - maximized for $\mathbf{y} = (1, 0, \ldots, 0)^\top$, maximum σ_1

Eckart–Young Theorem: revisted

- **Reduced rank SVD**:
  optimal low rank approximation in Frobenius norm
  - SVD of $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, define for $k \leq \text{rank}(\mathbf{A})$

$$\mathbf{A}_k := \sum_{i=1}^{k} \sigma_i \mathbf{u}_i \mathbf{v}_i^\top, \quad \text{rank}(\mathbf{A}_k) = k$$

  - then $\mathbf{A}_k$ is best Frobenius norm approximation in the sense that

$$\min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2 = \|\mathbf{A} - \mathbf{A_k}\|_F^2 = \sum_{r=k+1}^{\text{rank}(\mathbf{A})} \sigma_r^2$$

Spectral Norm Approximation

- $\mathbf{A}_k$ an optimal approximation in the sense of the spectral norm

$$\min_{\text{rank}(\mathbf{B})=k} \|\mathbf{A}-\mathbf{B}\|_2 = \|\mathbf{A}-\mathbf{A_k}\|_2 = \sigma_{k+1}$$

Section 3

SVD for Collaborative Filtering

SVD of Rating Matrix: Interpretation

$\mathbf{A}$ = rating matrix, then ...

- k dimensional ($k \leq \text{rank}(\mathbf{A})$) number of latent factors
- $\mathbf{U}$: users-to-factor association matrix
- $\mathbf{V}$: items-to-factor association matrix
- $\mathbf{D}$: level of strength of each factor

SVD For Collaborative Filtering

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top:$$

$$\begin{array}{c}
\begin{array}{ccccc} \text{Cremators} & \text{Evil spawn} & \text{Fatal justice} & \text{Clerks} & \text{American pie} \end{array} \\
\begin{pmatrix}
5 & 5 & 5 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
0 & 0 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 0 & 0 & 4 & 4
\end{pmatrix}
\end{array}
=
\begin{pmatrix}
0.57 & 0 & -0.80 & 0.06 & -0.04 & -0.06 & -0.04 \\
0.46 & 0 & 0.43 & 0.68 & -0.19 & -0.23 & -0.19 \\
0.57 & 0 & 0.37 & -0.70 & -0.08 & -0.11 & -0.08 \\
0.34 & 0 & 0.15 & 0.14 & 0.48 & 0.60 & 0.48 \\
0 & 0.52 & 0 & 0 & -0.71 & 0.35 & 0.28 \\
0 & 0.66 & 0 & 0 & 0.35 & -0.56 & 0.35 \\
0 & 0.52 & 0 & 0 & 0.28 & 0.35 & -0.71
\end{pmatrix}
\times$$

$$\begin{pmatrix}
15 & 0 & 0 & 0 & 0 \\
0 & 10.67 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0
\end{pmatrix}
\times
\begin{pmatrix}
0.57 & 0.57 & 0.57 & 0 & 0 \\
0 & 0 & 0 & 0.70 & 0.70 \\
0.81 & -0.40 & -0.40 & 0 & 0 \\
0 & 0.70 & -0.70 & 0 & 0 \\
0 & 0 & 0 & 0.70 & -0.70
\end{pmatrix}$$

SVD For Collaborative Filtering

Factors: **Horror, Comedy**

$\mathbf{U}$: users-to-factors association matrix.

$$\begin{array}{c} \begin{array}{ccccc} \text{Cremators} & \text{Evil spawn} & \text{Fatal justice} & \text{Clerks} & \text{American pie} \end{array} \\ \begin{pmatrix} 5 & 5 & 5 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 4 & 4 \end{pmatrix} \end{array} = \begin{array}{c} \begin{array}{cc} \text{Horror} & \text{Comedy} \end{array} \\ \begin{pmatrix} 0.57 & 0 \\ 0.46 & 0 \\ 0.57 & 0 \\ 0.34 & 0 \\ 0 & 0.52 \\ 0 & 0.66 \\ 0 & 0.52 \end{pmatrix} \end{array} \times \begin{pmatrix} 15 & 0 \\ 0 & 10.67 \end{pmatrix} \times \begin{pmatrix} 0.57 & 0.57 & 0.57 & 0 & 0 \\ 0 & 0 & 0 & 0.70 & 0.70 \end{pmatrix}$$

Q: What is the affinity between user1 and horror? 0.57

SVD For Collaborative Filtering

Factors: Horror, Comedy

$\mathbf{D}$: weight of different factors in the data.

Strength of Horror

Strength of Comedy

$$\begin{array}{c} \begin{array}{ccccc} \text{Cremators} & \text{Evil spawn} & \text{Fatal justice} & \text{Clerks} & \text{American pie} \end{array} \\ \begin{pmatrix} 5 & 5 & 5 & 0 & 0 \\ 4 & 4 & 4 & 0 & 0 \\ 5 & 5 & 5 & 0 & 0 \\ 3 & 3 & 3 & 0 & 0 \\ 0 & 0 & 0 & 4 & 4 \\ 0 & 0 & 0 & 5 & 5 \\ 0 & 0 & 0 & 4 & 4 \end{pmatrix} \end{array} = \begin{pmatrix} 0.57 & 0 \\ 0.46 & 0 \\ 0.57 & 0 \\ 0.34 & 0 \\ 0 & 0.52 \\ 0 & 0.66 \\ 0 & 0.52 \end{pmatrix} \times \begin{pmatrix} 15 & 0 \\ 0 & 10.67 \end{pmatrix} \times \begin{pmatrix} 0.57 & 0.57 & 0.57 & 0 & 0 \\ 0 & 0 & 0 & 0.70 & 0.70 \end{pmatrix}$$

Q: What is the expression of the comedy concept in the data? 10.67

SVD For Collaborative Filtering

Factors: Horror, Comedy

$\mathbf{V}$: Movies-to-factor association matrix.

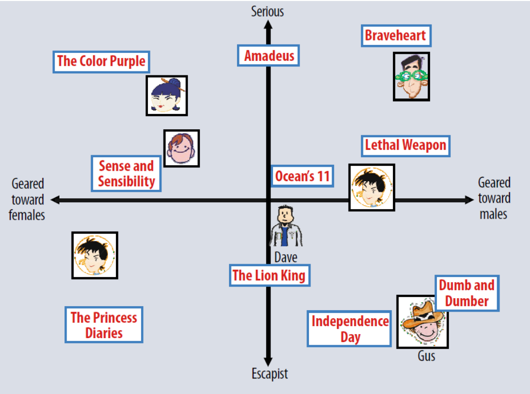
$$\begin{array}{c}
\begin{array}{ccccc} \text{Cremators} & \text{Evil spawn} & \text{Fatal justice} & \text{Clerks} & \text{American pie} \end{array} \\
\begin{pmatrix}
5 & 5 & 5 & 0 & 0 \\
4 & 4 & 4 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
3 & 3 & 3 & 0 & 0 \\
0 & 0 & 0 & 4 & 4 \\
0 & 0 & 0 & 5 & 5 \\
0 & 0 & 0 & 4 & 4
\end{pmatrix}
\end{array}
=
\begin{pmatrix}
0.57 & 0 \\
0.46 & 0 \\
0.57 & 0 \\
0.34 & 0 \\
0 & 0.52 \\
0 & 0.66 \\
0 & 0.52
\end{pmatrix}
\times
\begin{pmatrix}
15 & 0 \\
0 & 10.67
\end{pmatrix}
\times
\begin{pmatrix}
0.57 & 0.57 & 0.57 & 0 & 0 \\
0 & 0 & 0 & 0.70 & 0.70
\end{pmatrix}$$

Q: What is the similarity between Clerks and Horror? 0

What is the similarity between Clerks and Comedy? 0.7

Collaborative Filtering Example II

Characterization of the users and movies using two axes - male vs. female and serious vs. escapist.



* Ref: "Matrix factorization techniques for recommender systems"

http://www2.research.att.com/∼volinsky/papers/ieeecomputer.pdf.

Section 4

Alternating Least Squares

Beyond Singular Value Decomposition

- Is SVD the final answer for (low-rank) matrix decomposition?
  - **Eckart-Young theorem** guarantees:

$$\mathbf{A}_k = \operatorname*{arg\,min}_{\mathsf{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2$$

  - surprisingly: **not** a convex optimization problem!
  - convex combination of k-rank matrices is generally not rank k

$$\frac{1}{2}\underbrace{\begin{bmatrix}1 & 0\\ 0 & 0\end{bmatrix}}_{\text{rank } 1} + \frac{1}{2}\underbrace{\begin{bmatrix}0 & 0\\ 0 & 1\end{bmatrix}}_{\text{rank } 1} = \frac{1}{2}\underbrace{\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}}_{\text{rank } 2}$$

Beyond Singular Value Decomposition

- Problem: entries which are **unobserved** – not zero!
  - should optimize

$$\min_{\mathrm{rank}(\mathbf{B})=k} \left[\sum_{(i,j)\in\mathcal{I}} (a_{ij} - b_{ij})^2 \right], \quad \mathcal{I} = \{(i,j) : \text{observed}\}$$

  - instead of

$$\min_{\mathrm{rank}(\mathbf{B})=k} \left[\sum_{i,j} (a_{ij} - b_{ij})^2 \right] = \min_{\mathrm{rank}(\mathbf{B})=k} \|\mathbf{A} - \mathbf{B}\|_F^2$$

    - usually: mean zero $a_{ij} \leftarrow a_{ij} - \frac{1}{|\mathcal{I}|} \sum_{\mathcal{I}} a_{ij}$

Hardness of Matrix Reconstruction

- Define weighted Frobenius norm with regard to matrix $\mathbf{G} \geq \mathbf{0}$.

$$\|\mathbf{X}\|_{\mathbf{G}} := \sqrt{\sum_{i,j} g_{ij} x_{ij}^2}$$

  - special case: $g_{ij} \in \{0, 1\}$ (Boolean, partially observed matrix)

- Low-rank approximations are (in general) intrinsically hard

$$\mathbf{B}^* \xrightarrow{\min} \ell(\mathbf{B}) = \|\mathbf{A} - \mathbf{B}\|_{\mathbf{G}}^2, \quad \text{s.t. } \operatorname{rank}(\mathbf{B}) \leq k$$

  - is NP-hard (Gillis & Glineur, 2011) even for $k = 1$.
  - ... also holds for approximations with prescribed accuracy
  - ... also holds for binary $\mathbf{G}$

Matrix Factorization: Non-Convex Problem

- Singular Value Decomposition is not enough!
- **Non-convex** optimization problem
  - **variant A**: non-convex domain

    minimize convex objective over domain $\mathcal{Q}_k := \{\mathbf{B} : \text{rank}(\mathbf{B}) = k\}$
  - **variant B**: non-convex objective

    re-parametrize $\mathbf{B} = \mathbf{U}\mathbf{V}, \quad \mathbf{U} \in \mathbb{R}^{m \times k}, \mathbf{V} \in \mathbb{R}^{k \times n}$

    then $\text{rank}(\mathbf{B}) \leq k$ by definition

    e.g. $f(u,v) = (a - uv)^2,\ u_1 v_1 = u_2 v_2 = a \ \wedge \ u_1 \neq u_2$

    $$\Longrightarrow f(u_1, v_1) = f(u_2, v_2) = 0 \wedge f\left(\frac{u_1 + u_2}{2}, \frac{v_1 + v_2}{2}\right) > 0$$

Alternating Minimization

- Is there something **convex** about the **non-convex** objective?

$$f(\mathbf{U}, \mathbf{V}) = \frac{1}{|\mathcal{I}|} \sum_{(i,j) \in \mathcal{I}} (a_{ij} - \langle \mathbf{u}_i, \mathbf{v}_j \rangle)^2$$

  - for fixed $\mathbf{U}$, f is convex in $\mathbf{V}$ – for fixed $\mathbf{V}$, f is convex in $\mathbf{U}$
  - ... which does not mean f is jointly convex in $\mathbf{U}$ and $\mathbf{V}$

- Idea: perform **alternating minimization**

$$\mathbf{U} \leftarrow \arg\min_{\mathbf{U}} f(\mathbf{U}, \mathbf{V})$$

$$\mathbf{V} \leftarrow \arg\min_{\mathbf{V}} f(\mathbf{U}, \mathbf{V}), \quad \text{repeat until convergence}$$

  - f is never increased and lower bounded by 0

Alternating Least Squares

- Alternating minimization for low-rank matrix factorization = **alternating least squares**
  - decompose f into subproblems for columns of $\mathbf{V}$

$$f(\mathbf{U}, \mathbf{V}) = \sum_i \underbrace{\left[\sum_{j:(i,j)\in\mathcal{I}} (a_{ij} - \langle \mathbf{u}_j, \mathbf{v}_i \rangle)^2 \right]}_{=: f(\mathbf{U}, \mathbf{v}_i)}$$

  - least squares problem $f(\mathbf{U}, \mathbf{v}_i)$ for column $\mathbf{v}_i$ of $\mathbf{V}$
    - each of which can be solved independently!
  - by symmetry: also holds for $\mathbf{U} \leftrightarrow \mathbf{V}$

Frobenius Norm Regularization

- Typically: regularize matrix factors $\mathbf{U}, \mathbf{V}$
- (squared) Frobenius norm regularizer

$$\Omega(\mathbf{U}, \mathbf{V}) = \|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2$$

  - then

$$\text{minimize} \;\rightarrow\; f(\mathbf{U}, \mathbf{V}) + \mu\, \Omega(\mathbf{U}, \mathbf{V}), \quad \mu > 0$$

  - does not change separability structure of problem

ALS for Collaborative Filtering

- given low-dimensional representations for items
- compute for each user independently the best representation

- given low-dimensional representations for users
- compute for each item independently the best representation

- all optimization problems are least-square problems of small dimension

Section 5

Convex Relaxation

Nuclear Norm

- **Nuclear norm**

$$\|\mathbf{A}\|_* = \sum_i \sigma_i, \quad \sigma_i : \text{singular values of } \mathbf{A}$$

- Compare with Frobenius norm $\|\mathbf{A}\|_F = \sqrt{\sum_i \sigma_i^2}$
- Or, alternatively, if we define $\boldsymbol{\sigma}(\mathbf{A}) = (\sigma_1, \ldots, \sigma_n)$, then

$$\|\mathbf{A}\|_F = \|\boldsymbol{\sigma}(\mathbf{A})\|_{\mathbf{2}} \quad \text{whereas} \quad \|\mathbf{A}\|_* = \|\boldsymbol{\sigma}(\mathbf{A})\|_{\mathbf{1}}$$

- For a diagonal matrix $\mathbf{D}$, $\|\mathbf{D}\|_* = \mathrm{Tr}(\mathbf{D})$.

Nuclear Norm Minimization

- Exact reconstruction (Boolean $\mathbf{G}$)

$$\min_{\mathbf{B}} \|\mathbf{B}\|_* \quad \text{subject to} \quad \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}} = 0$$

- Approximate reconstruction

$$\min_{\mathbf{B}} \|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}}^2, \quad \text{s.t.} \quad \|\mathbf{B}\|_* \leq r$$

- Lagrangian formulation

$$\min_{\mathbf{B}} \left[\tfrac{1}{2\tau}\|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}}^2 + \|\mathbf{B}\|_*\right]$$

Nuclear Norm vs. Rank

- How does this relate to low rank approximation?
- Lower bound

$$\text{rank}(\mathbf{B}) \geq \|\mathbf{B}\|_*, \quad \text{for} \quad \|\mathbf{B}\|_2 \leq 1$$

  - in fact: tightest convex lower bound (Fazel 2002)

- **Convex relaxation**

$$\min_{\mathbf{B} \in \mathcal{P}_k} \|\mathbf{A} - \mathbf{B}\|^2_{\mathbf{G}}, \quad \mathcal{P}_k := \{\mathbf{B} : \|\mathbf{B}\|_* \leq k\}$$

  where

$$\mathcal{P}_k \supseteq \mathcal{Q}_k = \{\mathbf{B} : \text{rank}(\mathbf{B}) \leq k\}$$

SVD Thresholding

- How to solve optimization problems involving the nuclear norm?
- Fundamental result (due to Cai, Candès & Shen, 2008)

$$\mathbf{B}^* = \mathsf{shrink}_\tau(\mathbf{A}) := \arg\min_{\mathbf{B}} \left\{ \frac{1}{2}\|\mathbf{A}-\mathbf{B}\|_F^2 + \tau\|\mathbf{B}\|_* \right\}$$

  then with SVD $\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, $\mathbf{D} = \mathsf{diag}(\sigma_i)$, it holds that

$$\mathbf{B}^* = \mathbf{U}\mathbf{D}_\tau\mathbf{V}^\top, \quad \mathbf{D}_\tau = \mathsf{diag}\left(\max\{0, \sigma_i - \tau\}\right)$$

  - note: all singular values are reduced by at least τ

SVD Shrinkage Iterations

- SVD thresholding + projection = Shrinkage iterations (due to Cai, Candès & Shen, 2008)
- Define projection operator with regard to index set $\mathcal{I}$

$$\Pi(\mathbf{X}) = \begin{cases} x_{ij} & (i,j) \in \mathcal{I} \\ 0 & \text{otherwise} \end{cases}$$

- Iterative algorithm, initialized with $\mathbf{B}_0 = \mathbf{0}$

$$\mathbf{B}_{t+1} = \mathbf{B}_t + \eta_t \, \Pi(\mathbf{A} - \mathsf{shrink}_\tau(\mathbf{B}_t))$$

  - $\eta_t > 0$: learning rate sequence

SVD Shrinkage Iterations: Analysis

- $\mathbf{B}_t$ is a sequence of sparse matrices (efficiency!)
- It can be shown that[1] $\lim_{t\to\infty} \text{shrink}_\tau(\mathbf{B}_t) = \mathbf{B}^*$, the minimizer of

$$\mathbf{B}^* = \arg\min_{\mathbf{B}} \left\{ \|\mathbf{B}\|_* + \frac{1}{2\tau}\|\mathbf{B}\|_F^2 \right\}, \quad \text{s.t. } \Pi(\mathbf{A}-\mathbf{B}) = \mathbf{0}$$

- For large enough τ one finds a minimal nuclear-norm approximation to $\mathbf{A}$ that agrees on all observed entires.
- Can be extented to $\|\mathbf{A}-\mathbf{B}\|_{\mathbf{G}}$ residuals (by modifying Π)

[1] Upon appropriate choice of step sizes.

Exact Matrix Recovery

- Can use SVD-shrinkage iterations to solve convex relaxations.
- But: can we get any "generalization" guarantees ($\Pi(\mathbf{A}^*) = \mathbf{A}$)?

$$\mathbf{B}^* = \arg\min_{\mathbf{B}} \{\|\mathbf{B}\|_*\}, \quad \text{s.t. } \Pi(\mathbf{A} - \mathbf{B}) = \mathbf{0}$$

  - suprising (deep) result: **yes!**

- **Theorem**: Exact reconstruction of rank k matrix $\mathbf{A}^*$ w.h.p., if it is strongly incoherent (parameter μ, spread of singular values), if

$$|\mathcal{I}| \geq C\mu^4 k^2 n (\log n)^2 \in \tilde{\mathbf{O}}(n), \quad \text{typically} \quad \mu = \mathbf{O}(\sqrt{\log n})$$

  - due to Candes & Tao, 2010
  - explains, why $\|\cdot\|_*$ minimization works well in practice!