ECE 18-898G: Special Topics in Signal Processing: Sparsity, Structure, and Inference

Low-rank matrix recovery via nonconvex optimization

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

Spring 2018

Outline

- Low-rank matrix completion and recovery
- Nuclear norm minimization (last lecture)
  - RIP and low-rank matrix recovery
  - Matrix completion
  - Algorithms for nuclear norm minimization
- Non-convex methods (this lecture)
  - Global landscape
  - Spectral methods
  - (Projected) gradient descent

Why nonconvex?

- Consider completing an $n \times n$ matrix, with rank r:

$$\text{minimize}_{\boldsymbol{X}} \quad \|\mathcal{P}_\Omega(\boldsymbol{X} - \boldsymbol{M})\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \text{rank}(\boldsymbol{X}) \le r,$$

  where $r \ll n$.
  - The size of observation $|\Omega|$ is about $nr\mathrm{polylog}n$;
  - The degrees of freedom in $\boldsymbol{X}$ is about nr;

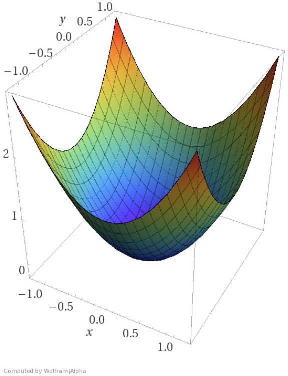
- Question: Can we develop algorithms that work with computational and memory complexity that nearly linear in n?

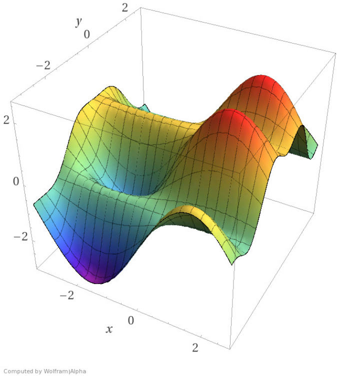
- This means that we don't even want to store the matrix $\boldsymbol{X}$ which takes n^2 storage.

- A nonconvex approach will store and update a "low-dimensional" representation of $\boldsymbol{X}$ throughout the execution of the algorithm.

Convex vs. nonconvex

$$\text{minimize}_{\boldsymbol{x}}\ f(\boldsymbol{x})$$





convex vs. **nonconvex**

Prelude: low-rank matrix approximation — an optimization perspective

Low-rank matrix approximation / PCA

Given $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ (not necessarily low-rank), solve the *low-rank approximation problem* (best rank-r approximation):

$$\widehat{\boldsymbol{M}} = \mathrm{argmin}_{\boldsymbol{X}} \ \|\boldsymbol{X} - \boldsymbol{M}\|_{\mathrm{F}}^2 \quad \text{s.t.} \quad \mathrm{rank}(\boldsymbol{X}) \leq r.$$

this is a nonconvex optimization problem.

The solution is known as the **Eckart-Young theorem**:

- denote the SVD of $\boldsymbol{M} = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$, where σ_i's are in a descending order; then

$$\widehat{\boldsymbol{M}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^\top.$$

nonconvex, but tractable.

Optimization viewpoint

Let us factorize $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{R}^{n\times r}$. Our problem is equivalent to

$$\text{minimize}_{\boldsymbol{U},\boldsymbol{V}} \; f(\boldsymbol{U}, \boldsymbol{V}) := \|\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2.$$

- The size of $\boldsymbol{U}, \boldsymbol{V}$ are of $O(nr)$, which is much smaller than $\boldsymbol{X}$;
- Identifiability issues: for any orthonormal $\boldsymbol{R} \in \mathbb{R}^{r\times r}$, we have

$$\boldsymbol{U}\boldsymbol{V}^\top = (\alpha \boldsymbol{U}\boldsymbol{R})(\alpha^{-1}\boldsymbol{V}\boldsymbol{R})^\top.$$

If $(\boldsymbol{U}, \boldsymbol{V})$ is a global minimizer (..), so does $(\alpha\boldsymbol{U}\boldsymbol{R}, \alpha^{-1}\boldsymbol{V}\boldsymbol{R})$.

Question: what does $f(\boldsymbol{U}, \boldsymbol{V})$ look like (landscape)? (we already found its global minima.)

The PSD case

For simplicity, consider the PSD case.

- Let $\boldsymbol{M}$ be PSD, so that $\boldsymbol{M} = \sum_{i=1}^{n} \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top$.
- Let $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{U}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{n\times r}$.

We're interested in the landscape of

$$f(\boldsymbol{U}) := \frac{1}{4}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2.$$

Identifiability: for any orthonormal $\boldsymbol{R} \in \mathbb{R}^{r\times r}$, we have

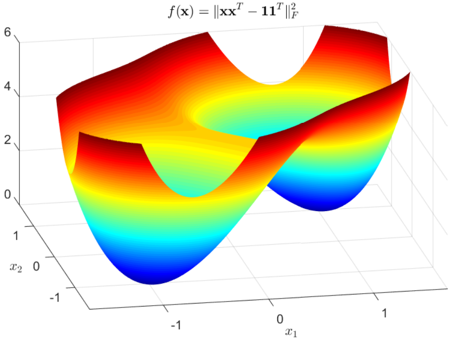
$$\boldsymbol{U}\boldsymbol{U}^\top = (\boldsymbol{U}\boldsymbol{R})(\boldsymbol{U}\boldsymbol{R})^\top.$$

make the exposition even simpler: set $r = 1$.

$$f(\boldsymbol{u}) = \frac{1}{4}\|\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2.$$

Good news: benign landscape

Take $f(\boldsymbol{u}) = \left\| \boldsymbol{u}\boldsymbol{u}^\top - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \right\|_{\mathrm{F}}^2$.



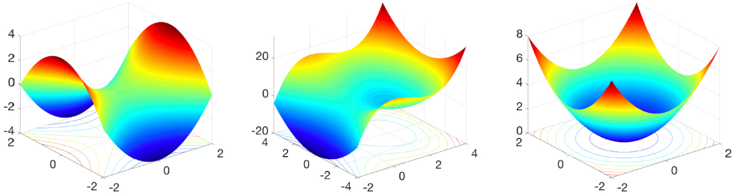
Global optima: $\boldsymbol{x} = \pm \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, strict saddle $\boldsymbol{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$. No "spurious" local minima.

Critical points

Definition 7.1

A first-order critical point (stationary point) satisfies

$$\nabla f(\boldsymbol{u}) = 0.$$



(a) strict saddle (b) local minimum (c) global minimum

Figure credit: Li et al., 2016

Critical points of $f(\boldsymbol{u})$

Any $\boldsymbol{u} \in \mathbb{R}^n$ satisfies

$$\nabla f(\boldsymbol{u}) = (\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{M})\boldsymbol{u} = 0.$$

$$\Updownarrow$$

$$\boldsymbol{M}\boldsymbol{u} = \|\boldsymbol{u}\|_2^2 \boldsymbol{u}$$

$$\Updownarrow$$

$\boldsymbol{u}$ aligns with eigenvectors of $\boldsymbol{M}$.

or

$$\boldsymbol{u} = \boldsymbol{0}.$$

Critical points of $f(\boldsymbol{u})$

Any $\boldsymbol{u} \in \mathbb{R}^n$ satisfies

$$\nabla f(\boldsymbol{u}) = (\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{M})\boldsymbol{u} = 0.$$

$$\Updownarrow$$

$$\boldsymbol{M}\boldsymbol{u} = \|\boldsymbol{u}\|_2^2 \boldsymbol{u}$$

$$\Updownarrow$$

$\boldsymbol{u}$ aligns with eigenvectors of $\boldsymbol{M}$.

or

$$\boldsymbol{u} = \boldsymbol{0}.$$

Since $\boldsymbol{M}\boldsymbol{u}_i = \sigma_i \boldsymbol{u}_i$, the set of critical points are given as

$$\{\sqrt{\sigma_i}\boldsymbol{u}_i, i = 1, \ldots, n\}.$$

Categorization of critical points

Need to examine the Hessian:

$$\nabla^2 f(\boldsymbol{u}) := 2\boldsymbol{u}\boldsymbol{u}^\top + \|\boldsymbol{u}\|_2^2 \boldsymbol{I} - \boldsymbol{M}.$$

- Plug in the non-zero critical points: $\tilde{\boldsymbol{u}}_k := \sqrt{\sigma_k}\boldsymbol{u}_k$,

$$\begin{aligned}
\nabla^2 f(\tilde{\boldsymbol{u}}_k) &= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \boldsymbol{I} - \boldsymbol{M} \\
&= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \left(\sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top \right) - \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \\
&= \sum_{i \neq k} (\sigma_k - \sigma_i) \boldsymbol{u}_i \boldsymbol{u}_i^\top + 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top
\end{aligned}$$

- Assume $\sigma_1 > \sigma_2 > \ldots > \sigma_n > 0$:
  - $k = 1$: $\nabla^2 f(\tilde{\boldsymbol{u}}_1) \succ 0 \rightarrow$ local minima
  - $1 < k \leq n$: $\lambda_{\min}(\nabla^2 f(\tilde{\boldsymbol{u}}_k)) < 0$, $\lambda_{\max}(\nabla^2 f(\tilde{\boldsymbol{u}}_k)) > 0$, $\rightarrow$ strict saddle
  - $\boldsymbol{u} = 0$: $\nabla^2 f(0) \preceq 0 \rightarrow$ local maxima

Categorization of critical points

Need to examine the Hessian:

$$\nabla^2 f(\boldsymbol{u}) := 2\boldsymbol{u}\boldsymbol{u}^\top + \|\boldsymbol{u}\|_2^2 \boldsymbol{I} - \boldsymbol{M}.$$

- Plug in the non-zero critical points: $\tilde{\boldsymbol{u}}_k := \sqrt{\sigma_k}\boldsymbol{u}_k$,

$$\begin{aligned}\nabla^2 f(\tilde{\boldsymbol{u}}_k) &= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \boldsymbol{I} - \boldsymbol{M} \\ &= 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top + \sigma_k \left(\sum_{i=1}^n \boldsymbol{u}_i \boldsymbol{u}_i^\top\right) - \sum_{i=1}^n \sigma_i \boldsymbol{u}_i \boldsymbol{u}_i^\top \\ &= \sum_{i \neq k} (\sigma_k - \sigma_i) \boldsymbol{u}_i \boldsymbol{u}_i^\top + 2\sigma_k \boldsymbol{u}_k \boldsymbol{u}_k^\top\end{aligned}$$

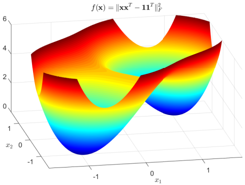
- Assume $\sigma_1 > \sigma_2 \geq \ldots \geq \sigma_n \geq 0$:
  - $k = 1$: $\nabla^2 f(\tilde{\boldsymbol{u}}_1) \succ 0 \rightarrow$ local minima
  - $1 < k \leq n$: $\lambda_{\min}(\nabla^2 f(\tilde{\boldsymbol{u}}_k)) < 0$, $\lambda_{\max}(\nabla^2 f(\tilde{\boldsymbol{u}}_k)) > 0$, $\rightarrow$ strict saddle
  - $\boldsymbol{u} = 0$: $\nabla^2 f(0) \prec 0 \rightarrow$ strict saddle

Summary

$$f(\boldsymbol{U}) := \frac{1}{4}\|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2, \qquad \boldsymbol{U} \in \mathbb{R}^{n\times r},$$

If $\sigma_r > \sigma_{r+1}$,

- **all local minima are global:** $\boldsymbol{U}$ contains the top-r eigenvectors up to an orthonormal transformation;
- **strict saddle points:** all stationary points are saddle points except the global optimum.



Undersampled regime

Consider linear measurements:

$$\boldsymbol{y} = \mathcal{A}(\boldsymbol{M}), \quad \boldsymbol{y} \in \mathbb{R}^m, \quad m \ll n^2$$

where $\boldsymbol{M} = \boldsymbol{U}_0 \boldsymbol{U}_0^\top \in \mathbb{R}^{n \times n}$ is **rank**-r, $r \ll n$, and PSD (for simplicity).

- The loss function we consider:

$$f(\boldsymbol{U}) := \frac{1}{4} \|\mathcal{A}(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M})\|_{\mathrm{F}}^2.$$

- If $\mathbb{E}[\mathcal{A}^*\mathcal{A}] = \mathcal{I}$, then

$$\mathbb{E}[f(\boldsymbol{U})] = \frac{1}{4} \|\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M}\|_{\mathrm{F}}^2.$$

- Does $f(\boldsymbol{U})$ inherit the benign landscape?

Landscape preserving under RIP

Recall the definition of RIP:

Definition 7.2

The rank-r restricted isometry constants δ_r is the smallest quantity

$$(1-\delta_r)\|\boldsymbol{X}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{X})\|_{\mathsf{F}}^2 \le (1+\delta_r)\|\boldsymbol{X}\|_{\mathsf{F}}^2, \qquad \forall \boldsymbol{X} : \mathsf{rank}(\boldsymbol{X}) \le r$$

Landscape preserving under RIP

Recall the definition of RIP:

Definition 7.2

The rank-r restricted isometry constants δ_r is the smallest quantity

$$(1-\delta_r)\|\boldsymbol{X}\|_{\mathsf{F}}^2 \le \|\mathcal{A}(\boldsymbol{X})\|_{\mathsf{F}}^2 \le (1+\delta_r)\|\boldsymbol{X}\|_{\mathsf{F}}^2, \qquad \forall \boldsymbol{X} : \mathsf{rank}(\boldsymbol{X}) \le r$$

Theorem 7.3 (Bhojanapalli et al.' 2016, Ge et al.' 2017)

If $\mathcal{A}$ satisfies the RIP with $\delta_{2r} < \frac{1}{10}$, then $f(\boldsymbol{U})$ satisfies

- *all local min are global: for any local minimum $\boldsymbol{U}$ of $f(\boldsymbol{U})$, it satisfies $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{M}$;*
- *strict saddle points: for non-local min critical point $\boldsymbol{U}$, it satisfies $\lambda_{\min}[\nabla^2 f(\boldsymbol{U})] \le -\frac{2}{5}\sigma_r$.*

Proof of Theorem 7.3 when $r = 1$

Without loss of generality, assume $\boldsymbol{M} = \boldsymbol{u}_0 \boldsymbol{u}_0^\top$, and $\sigma_1 = 1$.

- Step 1: check all the critical points:

$$\nabla f(\boldsymbol{u}) = \sum_{i=1}^{m} \langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{u}^\top - \underbrace{\boldsymbol{u}_0 \boldsymbol{u}_0^\top}_{\boldsymbol{M}} \rangle \boldsymbol{A}_i \boldsymbol{u} = 0$$

- Step 2: verify the Hessian at all the critical points:

$$\nabla^2 f(\boldsymbol{u}) = \sum_{i=1}^{m} \langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0 \boldsymbol{u}_0^\top \rangle \boldsymbol{A}_i + 2\boldsymbol{A}_i \boldsymbol{u}\boldsymbol{u}^\top \boldsymbol{A}_i^\top$$

Proof of Theorem 7.3 when $r = 1$

Proof: Assume $\boldsymbol{u}$ is first-order optimal. Consider the descent direction: $\boldsymbol{\Delta} = \boldsymbol{u} - \boldsymbol{u}_0$:

$$\boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{u}) \boldsymbol{\Delta} = \sum_{i=1}^m \left[\langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top \rangle \langle \boldsymbol{A}_i, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \rangle + 2\langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{\Delta}^\top \rangle^2 \right]$$

$$= \sum_{i=1}^m \left[\langle \boldsymbol{A}_i, \boldsymbol{\Delta}\boldsymbol{\Delta}^\top \rangle^2 - 3\langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top \rangle^2 \right].$$

where we have used the first order optimality condition.

Proof of Theorem 7.3 when $r = 1$

By the RIP property:

$$\boldsymbol{\Delta}^\top \nabla^2 f(\boldsymbol{u})\boldsymbol{\Delta} = \sum_{i=1}^{m} \left[\langle \boldsymbol{A}_i, (\boldsymbol{u}-\boldsymbol{u}_0)(\boldsymbol{u}-\boldsymbol{u}_0)^\top \rangle^2 - 3\langle \boldsymbol{A}_i, \boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top \rangle^2 \right]$$
$$\le (1+\delta)\|(\boldsymbol{u}-\boldsymbol{u}_0)(\boldsymbol{u}-\boldsymbol{u}_0)^\top\|_F^2 - 3(1-\delta)\|\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top\|_F^2$$
$$\le [2(1+\delta) - 3(1-\delta)]\|\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top\|_F^2$$
$$\le -(1-5\delta)\|\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top\|_F^2$$

where we use

$$\|(\boldsymbol{u}-\boldsymbol{u}_0)(\boldsymbol{u}-\boldsymbol{u}_0)^\top\|_F^2 \le 2\|\boldsymbol{u}\boldsymbol{u}^\top - \boldsymbol{u}_0\boldsymbol{u}_0^\top\|_F^2.$$

Landscape without RIP

In matrix completion, we need to regularize the loss function by promoting **incoherent** solutions: set

$$Q(\boldsymbol{U}) = \sum_{i=1}^{m} (\|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2 - \alpha)_+^4$$

where α is some regularization parameter, and $z_+ = \max\{z, 0\}$.

Landscape without RIP

In matrix completion, we need to regularize the loss function by promoting **incoherent** solutions: set

$$Q(\boldsymbol{U}) = \sum_{i=1}^{m} (\|\boldsymbol{e}_i^\top \boldsymbol{U}\|_2 - \alpha)_+^4$$

where α is some regularization parameter, and $z_+ = \max\{z, 0\}$.

Consider the loss function

$$f(\boldsymbol{U}) = \frac{1}{p}\|\mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{U}^\top - \boldsymbol{M})\|_F^2 + \lambda Q(\boldsymbol{U})$$

where λ is a regularization parameter.

- adding $Q(\boldsymbol{U})$ doesn't affect the global optimizer if α is set properly.

MC doesn't have spurious local minima

Theorem 7.4 (Ge et al, 2016)

If $p \gtrsim \frac{\mu^4 r^6 \log n}{n}$, $\alpha^2 = \Theta(\frac{\mu r \sigma_1}{n})$ *and* $\lambda = \Theta(\frac{n}{\mu r})$*, then with probability at least* $1 - n^{-1}$*,*

- *all local min are global: for any local minimum* $\boldsymbol{U}$ *of* $f(\boldsymbol{U})$*, it satisfies* $\boldsymbol{U}\boldsymbol{U}^\top = \boldsymbol{M}$*;*
- *saddle points that are not local minima are strict saddle points.*

- saddle-point escaping algorithms can be used to guarantee convergence to local minima, which in our problem are global minima.
- active research area for constructing saddle-point escaping algorithms: (perturbed) gradient descent, trust-region methods, etc...

Spectral methods: a one-shot approach

Setup

- Consider $\boldsymbol{M} \in \mathbb{R}^{n \times n}$ (square case for simplicity)
- $\text{rank}(\boldsymbol{M}) = r \ll n$
- The thin Singular value decomposition (SVD) of $\boldsymbol{M}$:

$$\boldsymbol{M} = \underbrace{\boldsymbol{U\Sigma V}^\top}_{(2n-r)r \text{ degrees of freedom}} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i \boldsymbol{v}_i^T$$

where $\boldsymbol{\Sigma} = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ contain all singular values $\{\sigma_i\}$;
$\boldsymbol{U} := [\boldsymbol{u}_1, \cdots, \boldsymbol{u}_r]$, $\boldsymbol{V} := [\boldsymbol{v}_1, \cdots, \boldsymbol{v}_r]$ consist of singular vectors

Signal + noise

$$(i, j) \in \Omega \quad \text{independently with prob. } p$$

One can write observation $\mathcal{P}_\Omega(\boldsymbol{M})$ as

$$\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) = \underbrace{\boldsymbol{M}}_{\text{signal}} + \underbrace{\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}}_{\text{noise}}$$

- Noise has mean zero: $\mathbb{E}\left[\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})\right] = \boldsymbol{M}$

Low-rank denoising

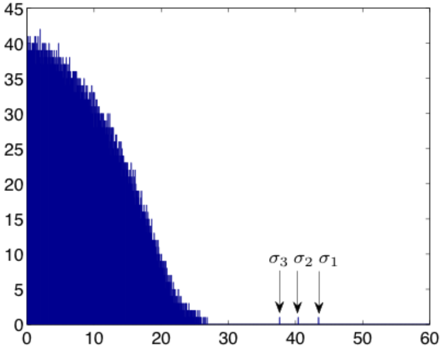
$$\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) = \underbrace{\boldsymbol{M}}_{\text{low-rank signal}} + \underbrace{\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}}_{:=\boldsymbol{E}\text{ (zero-mean noise)}}$$

Algorithm 7.1 Spectral method

$\hat{\boldsymbol{M}} \longleftarrow$ best rank-r approximation of $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})$

The spectral method can be solved via power methods or Lanczos methods, and we don't need to realize the matrix $\frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M})$.

Histograms of singular values of $\mathcal{P}_\Omega(\boldsymbol{M})$



A $10^4 \times 10^4$ random rank-3 matrix $\boldsymbol{M}$ with $p = 0.003$

Fig. credit: Keshavan, Montanari, Oh '10

Performance of spectral methods

Theorem 7.5 (Keshavan, Montanari, Oh '10)

Suppose number of observed entries m *obeys* $m \gtrsim n \log n$. *Then*

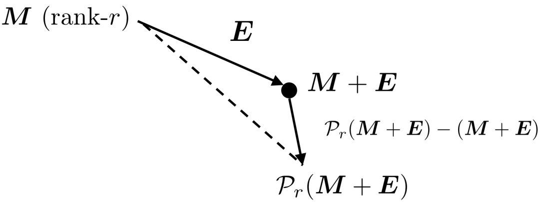
$$\frac{\|\widehat{\boldsymbol{M}} - \boldsymbol{M}\|_{\mathrm{F}}}{\|\boldsymbol{M}\|_{\mathrm{F}}} \lesssim \underbrace{\frac{\max_{i,j} |M_{i,j}|}{\frac{1}{n}\|\boldsymbol{M}\|_{\mathrm{F}}}}_{:=\nu} \cdot \sqrt{\frac{nr \log^2 n}{m}},$$

- ν reflects whether energy of $\boldsymbol{M}$ is spread out, $|M_{i,j}| \lesssim \mu r/n$;
- When $m \gg \nu^2 n \log^2 n$, estimate $\hat{\boldsymbol{M}}$ is very close to truth[1]
- Degrees of freedom $\asymp nr$

  $\longrightarrow$ nearly-optimal sample complexity for incoherent matrices

[1]The logarithmic factor can be improved.

Perturbation bounds



To ensure $\widehat{\boldsymbol{M}}$ is good estimate, it suffices to control noise $\boldsymbol{E}$

Lemma 7.6

Suppose rank$(\boldsymbol{M})=r$. *For any perturbation $\boldsymbol{E}$,*

$$\begin{aligned} \|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\| &\le 2\|\boldsymbol{E}\| \\ \|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\|_{\mathrm{F}} &\le 2\sqrt{2r}\|\boldsymbol{E}\| \end{aligned}$$

where $\mathcal{P}_r(\boldsymbol{X})$ is best rank-r approximation of $\boldsymbol{X}$.

Prior on matrix perturbation theory

Lemma 7.7 (Weyl's inequality, 1912)

Let $\boldsymbol{M}$, $\boldsymbol{E}$ be $n \times n$ matrices. Then

$$|\sigma_k(\boldsymbol{M} + \boldsymbol{E}) - \sigma_k(\boldsymbol{M})| \leq \|\boldsymbol{E}\|, \quad k = 1, \ldots, n.$$

Proof: Invoke the Courant-Fisher Minimax Characterization:

$$\sigma_k(\boldsymbol{A}) = \max_{\dim(\mathcal{S})=k} \min_{0 \neq \boldsymbol{v} \in \mathcal{S}} \frac{\|\boldsymbol{A}\boldsymbol{v}\|_2}{\|\boldsymbol{v}\|_2}.$$

Proof of Lemma 7.6

By matrix perturbation theory,

$$
\begin{aligned}
&\|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\| \\
&\overset{\text{triangle inequality}}{\leq} \quad \|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-(\boldsymbol{M}+\boldsymbol{E})\| + \|(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\| \\
&\leq \quad \sigma_{r+1}(\boldsymbol{M}+\boldsymbol{E}) + \|\boldsymbol{E}\| \\
&\overset{\text{Weyl's inequality}}{\leq} \quad \underbrace{\sigma_{r+1}(\boldsymbol{M})}_{=\boldsymbol{0}} + \|\boldsymbol{E}\| + \|\boldsymbol{E}\| \;=\; 2\|\boldsymbol{E}\|.
\end{aligned}
$$

The 2nd inequality of Lemma 7.6 follows since both $\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})$ and $\boldsymbol{M}$ are rank-r, and hence

$$\|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\|_{\mathrm{F}} \leq \sqrt{2r}\|\mathcal{P}_r(\boldsymbol{M}+\boldsymbol{E})-\boldsymbol{M}\|.$$

Controlling the noise

Recall that entries of $\boldsymbol{E} = \frac{1}{p}\mathcal{P}_\Omega(\boldsymbol{M}) - \boldsymbol{M}$ are zero-mean and independent.

A bit of random matrix theory ...

Lemma 7.8 (Chapter 2.3, Tao '12)

Suppose $\boldsymbol{X} \in \mathbb{R}^{n\times n}$ *is a random symmetric matrix obeying*

- $\{X_{i,j} : i < j\}$ *are independent*
- $\mathbb{E}[X_{i,j}] = 0$ *and* $\mathsf{Var}[X_{i,j}] \lesssim 1$
- $\max_{i,j} |X_{i,j}| \lesssim \sqrt{n}$

Then $\|\boldsymbol{X}\| \lesssim \sqrt{n}\log n$.

Proof of Theorem 7.5

If we look at the zero-mean matrix $\tilde{\boldsymbol{E}} = \frac{\sqrt{p}}{\mu/n}\boldsymbol{E}$, then

$$\begin{aligned} \mathsf{Var}\left[\tilde{E}_{i,j}\right] &= p(1-p)\cdot\left(\frac{\sqrt{p}}{\mu/n}\cdot\frac{1}{p}M_{i,j}\right)^2 \leq \left(\frac{M_{i,j}}{\mu/n}\right)^2 \lesssim 1, \\ |\tilde{E}_{i,j}| &\leq \frac{|M_{i,j}|}{\sqrt{p}\mu/n} \lesssim \frac{1}{\sqrt{p}}, \end{aligned}$$

where we have used the fact

$$|M_{i,j}| = \left|\boldsymbol{e}_i^\top \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{e}_j\right| \leq \|\boldsymbol{U}^\top \boldsymbol{e}_i\|\cdot\sigma_1\cdot\|\boldsymbol{V}^\top \boldsymbol{e}_j\| \overset{\text{(by our assumptions)}}{\lesssim} \frac{\mu r}{n} \asymp \frac{\mu}{n}$$

Lemma 7.8 tells us that if $p \gtrsim \frac{\log n}{n}$, then

$$\|\tilde{\boldsymbol{E}}\| \lesssim \sqrt{n}\log n \quad\Longleftrightarrow\quad \|\boldsymbol{E}\| \lesssim \frac{\mu}{\sqrt{pn}}\log n$$

This together with Lemma 7.6 and the fact $m \asymp pn^2$ establishes Theorem 7.5.

Gradient methods: iterative refinements

Iterative methods: an overview

$$\text{minimize}_{\boldsymbol{U},\boldsymbol{V}}\ f(\boldsymbol{U},\boldsymbol{V}) := \|\mathcal{P}_\Omega(\boldsymbol{U}\boldsymbol{V}^\top - \boldsymbol{M})\|_\mathrm{F}^2.$$

- **Gradient descent:** (our focus)

$$\boldsymbol{U}_{t+1} = \mathcal{P}_{\boldsymbol{U}}\Big[\boldsymbol{U}_t - \eta_t \nabla_{\boldsymbol{U}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)\Big],$$
$$\boldsymbol{V}_{t+1} = \mathcal{P}_{\boldsymbol{V}}\Big[\boldsymbol{V}_t - \eta_t \nabla_{\boldsymbol{V}} f(\boldsymbol{U}_t, \boldsymbol{V}_t)\Big].$$

  where η_t is the step size and $\mathcal{P}_{\boldsymbol{U}}$, $\mathcal{P}_{\boldsymbol{V}}$ denote the Euclidean projection onto some contraint sets;

- **Alternating minimization**: One optimizes $\boldsymbol{U}$, $\boldsymbol{V}$ alternatively while fixing the other, which is a convex problem.

$$\boldsymbol{U}_{t+1} = \text{argmin}_{\boldsymbol{U}}\ f(\boldsymbol{U}, \boldsymbol{V}_t),$$
$$\boldsymbol{V}_{t+1} = \text{argmin}_{\boldsymbol{V}}\ f(\boldsymbol{U}_{t+1}, \boldsymbol{V}).$$

Gradient descent for matrix completion

$$\text{minimize}_{\boldsymbol{X} \in \mathbb{R}^{n \times r}} \quad f(\boldsymbol{X}) = \sum_{(j,k) \in \Omega} \left(\boldsymbol{e}_j^\top \boldsymbol{X} \boldsymbol{X}^\top \boldsymbol{e}_k - M_{j,k} \right)^2$$

Algorithm 7.2 Gradient descent for MC

Input: $\boldsymbol{Y} = [Y_{j,k}]_{1 \le j,k \le n}$, r, p.

Spectral initialization: Let $\boldsymbol{U}^0 \boldsymbol{\Sigma}^0 \boldsymbol{U}^{0\top}$ be the rank-r eigendecomposition of

$$\boldsymbol{M}^0 := \frac{1}{p} \mathcal{P}_\Omega(\boldsymbol{Y}),$$

and set $\boldsymbol{X}^0 = \boldsymbol{U}^0 \left(\boldsymbol{\Sigma}^0 \right)^{1/2}$.

Gradient updates: **for** $t = 0, 1, 2, \ldots, T-1$ **do**

$$\boldsymbol{X}^{t+1} = \boldsymbol{X}^t - \eta_t \nabla f \left(\boldsymbol{X}^t \right).$$

Gradient descent for matrix completion

Define the optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\natural$ as

$$\boldsymbol{Q}^t := \text{argmin}_{\boldsymbol{R} \in \mathcal{O}^{r\times r}} \left\| \boldsymbol{X}^t \boldsymbol{R} - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}.$$

Theorem 7.9 (Ma et al., 2017)

Suppose $\boldsymbol{M} = \boldsymbol{X}^\natural \boldsymbol{X}^{\natural\top}$ is rank-r, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|_{\mathrm{F}}$,
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|\boldsymbol{X}^\natural\|$, *(spectral)*
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|\boldsymbol{X}^\natural\|_{2,\infty}$, *(incoherence)*

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$, if step size $\eta \asymp 1/\sigma_{max}$ and sample complexity $n^2 p \gtrsim \mu^3 n r^3 \log^3 n$.

Gradient descent for matrix completion

Define the optimal transform from the tth iterate $\boldsymbol{X}^t$ to $\boldsymbol{X}^\natural$ as

$$\boldsymbol{Q}^t := \mathrm{argmin}_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \left\|\boldsymbol{X}^t\boldsymbol{R} - \boldsymbol{X}^\natural\right\|_{\mathrm{F}}.$$

Theorem 7.9 (Ma et al., 2017)

Suppose $\boldsymbol{M} = \boldsymbol{X}^\natural\boldsymbol{X}^{\natural\top}$ is rank-r, incoherent and well-conditioned. Vanilla GD (with spectral initialization) achieves

- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\|_{\mathrm{F}} \lesssim \rho^t\mu r\frac{1}{\sqrt{np}}\|\boldsymbol{X}^\natural\|_{\mathrm{F}}$,
- $\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\| \lesssim \rho^t\mu r\frac{1}{\sqrt{np}}\|\boldsymbol{X}^\natural\|$, *(spectral)*
- $\left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{2,\infty} \lesssim \rho^t\mu r\sqrt{\frac{\log n}{np}}\|\boldsymbol{X}^\natural\|_{2,\infty}$, *(incoherence)*

where $\rho = 1 - \frac{\sigma_{\min}\eta}{5} < 1$, if step size $\eta \asymp 1/\sigma_{max}$ and sample complexity $n^2p \gtrsim \mu^3nr^3\log^3 n$.

- ***linear convergence*** of $\left\|\boldsymbol{X}^t\boldsymbol{X}^{t\top} - \boldsymbol{M}^\natural\right\|$ in Frobenius, spectral and infinity norms.

Numerical evidence for noiseless data

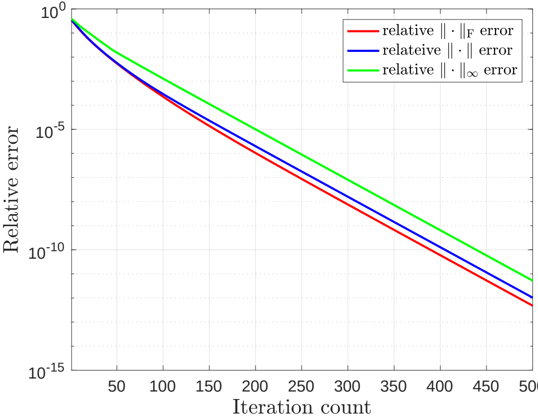


Figure 7.1: Relative error of $\boldsymbol{X}^t\boldsymbol{X}^{t\top}$ (measured by $\|\cdot\|_\mathrm{F}, \|\cdot\|, \|\cdot\|_\infty$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$.

Numerical evidence for noisy data

Set SNR $:= \frac{\|\boldsymbol{M}^\natural\|_\mathrm{F}^2}{n^2\sigma^2}$.

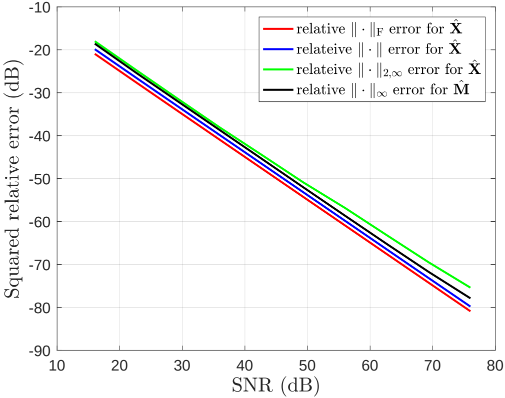


Figure 7.2: Squared relative error of the estimate $\hat{\boldsymbol{X}}$ (measured by $\|\cdot\|_\mathrm{F}, \|\cdot\|, \|\cdot\|_{2,\infty}$) and $\hat{\boldsymbol{M}} = \hat{\boldsymbol{X}}\hat{\boldsymbol{X}}^\top$ (measured by $\|\cdot\|_\infty$) vs. SNR, where $n = 500$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$.

Restricted strong convexity and smoothness

Lemma 7.10 (Restricted strong convexity and smoothness)

Suppose that $n^2 p \geq C\kappa^2 \mu r n \log n$ for some $C > 0$. Then with high probability, the Hessian $\nabla^2 f(\boldsymbol{X})$ obeys

$$\mathrm{vec}\left(\boldsymbol{V}\right)^\top \nabla^2 f\left(\boldsymbol{X}\right) \mathrm{vec}\left(\boldsymbol{V}\right) \geq \frac{\sigma_{\min}}{2} \|\boldsymbol{V}\|_{\mathrm{F}}^2 \quad \textit{(restricted strong convexity)}$$

$$\left\|\nabla^2 f\left(\boldsymbol{X}\right)\right\| \leq \frac{5}{2}\sigma_{\max} \quad \textit{(smoothness)}$$

for all $\boldsymbol{X}$ and $\boldsymbol{V} = \boldsymbol{Y}\boldsymbol{H}_Y - \boldsymbol{Z}$, $\boldsymbol{H}_Y := \arg\min_{\boldsymbol{R}\in\mathcal{O}^{r\times r}} \|\boldsymbol{Y}\boldsymbol{R} - \boldsymbol{Z}\|_{\mathrm{F}}$ satisfying

- $\left\|\boldsymbol{X} - \boldsymbol{X}^\natural\right\|_{2,\infty} \leq \epsilon \left\|\boldsymbol{X}^\natural\right\|_{2,\infty}$ *(incoherence region)*,
- $\|\boldsymbol{Z} - \boldsymbol{X}^\natural\| \leq \delta \|\boldsymbol{X}^\natural\|$,

where $\epsilon \ll 1/\sqrt{\kappa^3 \mu r \log^2 n}$ and $\delta \ll 1/\kappa$.

Linear convergence induction I

Given the definition of $\boldsymbol{Q}^{t+1}$, we have

$$\begin{aligned}
\left\|\boldsymbol{X}^{t+1}\boldsymbol{Q}^{t+1}-\boldsymbol{X}^{\natural}\right\|_{\mathrm{F}} &\leq \left\|\boldsymbol{X}^{t+1}\boldsymbol{Q}^{t}-\boldsymbol{X}^{\natural}\right\|_{\mathrm{F}} \\
&\overset{\text{(i)}}{=} \left\|\left[\boldsymbol{X}^{t}-\eta\nabla f\left(\boldsymbol{X}^{t}\right)\right]\boldsymbol{Q}^{t}-\boldsymbol{X}^{\natural}\right\|_{\mathrm{F}} \\
&\overset{\text{(ii)}}{=} \left\|\boldsymbol{X}^{t}\boldsymbol{Q}^{t}-\eta\nabla f(\boldsymbol{X}^{t}\boldsymbol{Q}^{t})-\boldsymbol{X}^{\natural}\right\|_{\mathrm{F}} \\
&\overset{\text{(iii)}}{=} \underbrace{\left\|\boldsymbol{X}^{t}\boldsymbol{Q}^{t}-\eta\nabla f(\boldsymbol{X}^{t}\boldsymbol{Q}^{t})-\left(\boldsymbol{X}^{\natural}-\eta\nabla f(\boldsymbol{X}^{\natural})\right)\right\|_{\mathrm{F}}}_{:=\alpha},
\end{aligned}$$

where (i) follows from the GD rule, (ii) follows from the identity $\nabla f(\boldsymbol{X}^{t}\boldsymbol{R}) = \nabla f\left(\boldsymbol{X}^{t}\right)\boldsymbol{R}$ for any $\boldsymbol{R}\in\mathcal{O}^{r\times r}$, and (iii) follows from $\nabla f(\boldsymbol{X}^{\natural}) = \boldsymbol{0}$.

Linear convergence induction II

The fundamental theorem of calculus reveals

$$
\begin{aligned}
&\mathrm{vec}\left[\boldsymbol{X}^t\boldsymbol{Q}^t - \eta\nabla f(\boldsymbol{X}^t\boldsymbol{Q}^t) - \left(\boldsymbol{X}^\natural - \eta\nabla f(\boldsymbol{X}^\natural)\right)\right] \\
&\quad = \mathrm{vec}\left[\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right] - \eta \cdot \mathrm{vec}\left[\nabla f(\boldsymbol{X}^t\boldsymbol{Q}^t) - \nabla f\left(\boldsymbol{X}^\natural\right)\right] \\
&\quad = \Bigg(\boldsymbol{I}_{nr} - \eta \underbrace{\int_0^1 \nabla^2 f\left(\boldsymbol{X}(\tau)\right) \mathrm{d}\tau}_{:=\boldsymbol{A}}\Bigg) \mathrm{vec}\left(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right), \qquad (7.1)
\end{aligned}
$$

where we denote $\boldsymbol{X}(\tau) := \boldsymbol{X}^\natural + \tau(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural)$. Taking the squared Euclidean norm of both sides of the equality (7.1) leads to

$$
\begin{aligned}
\alpha^2 &= \mathrm{vec}\big(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\big)^\top \left(\boldsymbol{I}_{nr} - \eta\boldsymbol{A}\right)^2 \mathrm{vec}\big(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\big) \\
&\le \left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{\mathrm{F}}^2 + \eta^2 \left\|\boldsymbol{A}\right\|^2 \left\|\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\right\|_{\mathrm{F}}^2 \\
&\qquad - 2\eta\, \mathrm{vec}\big(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\big)^\top \boldsymbol{A}\, \mathrm{vec}\big(\boldsymbol{X}^t\boldsymbol{Q}^t - \boldsymbol{X}^\natural\big), \qquad (7.2)
\end{aligned}
$$

Linear convergence induction III

Based on the incoherence of $\boldsymbol{X}^\natural$ and $\boldsymbol{X}^t$, $\forall \tau \in [0,1]$,

$$\left\| \boldsymbol{X}(\tau) - \boldsymbol{X}^\natural \right\|_{2,\infty} \le \underbrace{\left\| \boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural \right\|_{2,\infty} \le C\mu r \sqrt{\frac{\log n}{np}} \left\| \boldsymbol{X}^\natural \right\|_{2,\infty}}_{\text{incoherence hypothesis}}.$$

Taking $\boldsymbol{X} = \boldsymbol{X}(\tau), \boldsymbol{Y} = \boldsymbol{X}^t$ and $\boldsymbol{Z} = \boldsymbol{X}^\natural$ in Lemma 7.10, one can easily verify the assumptions therein given $n^2 p \gg \kappa^3 \mu^3 r^3 n \log^3 n$. Hence,

$$\mathrm{vec}\big(\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural\big)^\top \boldsymbol{A}\ \mathrm{vec}\big(\boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural\big) \ge \frac{\sigma_{\min}}{2} \left\| \boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}^2$$

and

$$\|\boldsymbol{A}\| \le \frac{5}{2}\sigma_{\max}.$$

Linear convergence induction IV

Substituting these two inequalities into (7.2) yields

$$\begin{aligned}\alpha^2 &\le \left(1 + \frac{25}{4}\eta^2\sigma_{\max}^2 - \sigma_{\min}\eta\right) \left\| \boldsymbol{X}^t \hat{\boldsymbol{H}}^t - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}^2 \\ &\le \left(1 - \frac{\sigma_{\min}}{2}\eta\right) \left\| \boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}^2\end{aligned}$$

as long as $0 < \eta \le (2\sigma_{\min})/(25\sigma_{\max}^2)$, which further implies that

$$\alpha \le \left(1 - \frac{\sigma_{\min}}{4}\eta\right) \left\| \boldsymbol{X}^t \boldsymbol{Q}^t - \boldsymbol{X}^\natural \right\|_{\mathrm{F}}.$$

The incoherence hypothesis is important for fast convergence: the fact that $\boldsymbol{X}^t$ stays incoherent throughout the execution is called "implicit regularization" and can be established by a leave-one-out analysis trick [Ma et al., 2017].

Reference

[1] "*Guaranteed matrix completion via non-convex factorization*," R. Sun, T. Luo, *IEEE Transactions on Information Theory*, 2016.

[2] "*The rotation of eigenvectors by a perturbation*," C. Davis, and W. Kahan, *SIAM Journal on Numerical Analysis*, 1970.

[3] "*Matrix completion from a few entries*," R. Keshavan, A. Montanari, and S. Oh, *IEEE Transactions on Information Theory*, 2010.

[4] "*Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees*," Y. Chen and M. Wainwright, *arXiv preprint arXiv:1509.03025*, 2015.

[5] "*Implicit Regularization in Nonconvex Statistical Estimation: Gradient Descent Converges Linearly for Phase Retrieval, Matrix Completion and Blind Deconvolution*," C. Ma, K. Wang, Y. Chi and Y. Chen, *arXiv preprint arXiv:1711.10467*, 2017.

Reference

[6] "*No Spurious Local Minima in Nonconvex Low Rank Problems: A Unified Geometric Analysis*," R. Ge, C. Jin, and Y. Zheng, *ICML*, 2017.

[7] "*Symmetry, Saddle Points, and Global Optimization Landscape of Nonconvex Matrix Factorization*," X. Li *et al.*, *arXiv preprint arxiv:1612.09296*, 2016.

[8] "*Topics in random matrix theory*," T. Tao, *American mathematical society*, 2012.

[9] "*Harnessing Structures in Big Data via Guaranteed Low-Rank Matrix Estimation*," Y. Chen, and Y. Chi, *arXiv preprint arXiv:1802.08397*, 2018.

[10] "*How to escape saddle points efficiently*," Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I, *arXiv preprint arXiv:1703.00887*, 2017.