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 $r$:

$$\text{minimize}_X \| P_{\Omega}(X - M) \|_F^2 \quad \text{s.t.} \quad \text{rank}(X) \leq r,$$

where $r \ll n$.

  ○ The size of observation $|\Omega|$ is about $nr \text{polylog} n$;  
  ○ The degrees of freedom in $X$ 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 $X$ which takes $n^2$ storage.

• A nonconvex approach will store and update a “low-dimensional” representation of $X$ throughout the execution of the algorithm.
Convex vs. nonconvex

\[
\text{minimize}_x \ f(x)
\]

convex \hspace{2cm} \text{vs.} \hspace{2cm} \text{nonconvex}
Prelude: low-rank matrix approximation — an optimization perspective
Given $M \in \mathbb{R}^{n \times n}$ (not necessarily low-rank), solve the low-rank approximation problem (best rank-$r$ approximation):

$$\hat{M} = \operatorname{argmin}_X \|X - M\|_F^2 \quad \text{s.t. } \operatorname{rank}(X) \leq r.$$ 

this is a nonconvex optimization problem.

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

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

$$\hat{M} = \sum_{i=1}^{r} \sigma_i u_i v_i^\top.$$ 

nonconvex, but tractable.
Optimization viewpoint

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

$$\text{minimize}_{U,V} \ f(U, V) := \|UV^\top - M\|_F^2.$$  

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

$$UV^\top = (\alpha UR)(\alpha^{-1} VR)^\top.$$  

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

**Question:** what does $f(U, V)$ look like (landscape)? (we already found its global minima.)
The PSD case

For simplicity, consider the PSD case.

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

We’re interested in the landscape of

$$f(U) := \frac{1}{4} \|UU^\top - M\|_F^2.$$

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

$$UU^\top = (UR)(UR)^\top.$$

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

$$f(u) = \frac{1}{4} \|uu^\top - M\|_F^2.$$
Good news: benign landscape

Take $f(u) = \| uu^T - \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \|_F^2$.

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

**Definition 7.1**

A first-order critical point (stationary point) satisfies

\[ \nabla f(u) = 0. \]

Figure credit: Li et al., 2016
Critical points of $f(u)$

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

$$\nabla f(u) = (uu^\top - M)u = 0.$$  

\[\iff\]

$$Mu = \|u\|_2^2 u$$

\[\iff\]

$u$ aligns with eigenvectors of $M$.

or

$$u = 0.$$
Critical points of $f(u)$

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

$$\nabla f(u) = (uu^\top - M)u = 0.$$  

$\Uparrow$

$$Mu = \|u\|^2_2 u$$  

$\Uparrow$

$u$ aligns with eigenvectors of $M$.

or

$$u = 0.$$  

Since $Mu_i = \sigma_i u_i$, the set of critical points are given as

$$\{\sqrt{\sigma_i}u_i, i = 1, \ldots, n\}.$$
Categorization of critical points

Need to examine the Hessian:

\[ \nabla^2 f(u) := 2uu^\top + \|u\|_2^2 I - M. \]

- Plug in the non-zero critical points: \( \tilde{u}_k := \sqrt{\sigma_k} u_k \),

\[
\nabla^2 f(\tilde{u}_k) = 2\sigma_k u_k u_k^\top + \sigma_k I - M
\]

\[
= 2\sigma_k u_k u_k^\top + \sigma_k \left( \sum_{i=1}^n u_i u_i^\top \right) - \sum_{i=1}^n \sigma_i u_i u_i^\top
\]

\[
= \sum_{i \neq k} (\sigma_k - \sigma_i) u_i u_i^\top + 2\sigma_k u_k u_k^\top
\]

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

Need to examine the Hessian:

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

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

  $$\nabla^2 f(\tilde{u}_k) = 2\sigma_ku_ku_k^\top + \sigma_kI - M$$

  $$= 2\sigma_ku_ku_k^\top + \sigma_k \left( \sum_{i=1}^n u_iu_i^\top \right) - \sum_{i=1}^n \sigma_iu_iu_i^\top$$

  $$= \sum_{i \neq k}(\sigma_k - \sigma_i)u_iu_i^\top + 2\sigma_ku_ku_k^\top$$

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

\[ f(U) := \frac{1}{4} \| UU^\top - M \|_F^2, \quad U \in \mathbb{R}^{n \times r}, \]

If \( \sigma_r > \sigma_{r+1} \),

- **all local minima are global:** \( U \) 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:

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

where \( M = U_0 U_0^\top \in \mathbb{R}^{n \times n} \) is rank-\( r \), \( r \ll n \), and PSD (for simplicity).

- The loss function we consider:

\[ f(U) := \frac{1}{4} \| \mathcal{A}(UU^\top - M) \|_F^2. \]

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

\[ \mathbb{E}[f(U)] = \frac{1}{4} \| UU^\top - M \|_F^2. \]

- Does \( f(U) \) inherit the benign landscape?
Landscape preserving under RIP

Recall the definition of RIP:

**Definition 7.2**

The rank-\( r \) restricted isometry constants \( \delta_r \) is the smallest quantity

\[
(1 - \delta_r) \|X\|_F^2 \leq \|A(X)\|_F^2 \leq (1 + \delta_r) \|X\|_F^2, \quad \forall X : \text{rank}(X) \leq r
\]
Recall the definition of RIP:

**Definition 7.2**

The rank-$r$ restricted isometry constants $\delta_r$ is the smallest quantity

$$(1 - \delta_r)\|X\|_F^2 \leq \|A(X)\|_F^2 \leq (1 + \delta_r)\|X\|_F^2, \quad \forall X : \text{rank}(X) \leq r$$

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

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

- all local min are global: for any local minimum $U$ of $f(U)$, it satisfies $UU^\top = M$;
- strict saddle points: for non-local min critical point $U$, it satisfies $\lambda_{\min}[\nabla^2 f(U)] \leq -\frac{2}{5}\sigma_r$. 
Proof of Theorem 7.3 when $r = 1$

Without loss of generality, assume $M = u_0 u_0^\top$, and $\sigma_1 = 1$.

- **Step 1**: check all the critical points:

  $$\nabla f(u) = \sum_{i=1}^{m} \left( \langle A_i, uu^\top - u_0 u_0^\top \rangle A_i + 2 A_i uu^\top A_i^\top \right) = 0$$

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

  $$\nabla^2 f(u) = \sum_{i=1}^{m} \left( \langle A_i, uu^\top - u_0 u_0^\top \rangle A_i + 2 A_i uu^\top A_i^\top \right)$$
Proof of Theorem 7.3 when \( r = 1 \)

Proof: Assume \( u \) is first-order optimal. Consider the descent direction: \( \Delta = u - u_0 \):

\[
\Delta^\top \nabla^2 f(u) \Delta = \sum_{i=1}^{m} \left[ \langle A_i, uu^\top - u_0 u_0^\top \rangle \langle A_i, \Delta \Delta^\top \rangle + 2 \langle A_i, u \Delta^\top \rangle^2 \right]
\]

\[
= \sum_{i=1}^{m} \left[ \langle A_i, \Delta \Delta^\top \rangle^2 - 3 \langle A_i, uu^\top - u_0 u_0^\top \rangle^2 \right].
\]

where we have used the first order optimality condition.
Proof of Theorem 7.3 when $r = 1$

By the RIP property:

$$\Delta^\top \nabla^2 f(u) \Delta = \sum_{i=1}^{m} \left[ \langle A_i, (u - u_0)(u - u_0)^\top \rangle^2 - 3\langle A_i, uu^\top - u_0u_0^\top \rangle^2 \right]$$

$$\leq (1 + \delta)\|(u - u_0)(u - u_0)^\top\|_F^2 - 3(1 - \delta)\|uu^\top - u_0u_0^\top\|_F^2$$

$$\leq [2(1 + \delta) - 3(1 - \delta)]\|uu^\top - u_0u_0^\top\|_F^2$$

$$\leq -(1 - 5\delta)\|uu^\top - u_0u_0^\top\|_F^2$$

where we use

$$\|(u - u_0)(u - u_0)^\top\|_F^2 \leq 2\|uu^\top - u_0u_0^\top\|_F^2.$$
In matrix completion, we need to regularize the loss function by promoting **incoherent** solutions: set

\[
Q(U) = \sum_{i=1}^{m} (\| e_i^T U \|_2 - \alpha)_+^4
\]

where \( \alpha \) is some regularization parameter, and \( z_+ = \max\{z, 0\} \).
In matrix completion, we need to regularize the loss function by promoting \textbf{incoherent} solutions: set

\[
Q(U) = \sum_{i=1}^{m} (\| e_i^\top U \|_2^2 - \alpha)_+^4
\]

where \(\alpha\) is some regularization parameter, and \(z_+ = \max\{z, 0\}\).

Consider the loss function

\[
f(U) = \frac{1}{p} \| P_\Omega(UU^\top - M) \|_F^2 + \lambda Q(U)
\]

where \(\lambda\) is a regularization parameter.

- adding \(Q(U)\) doesn’t affect the global optimizer if \(\alpha\) is set properly.
MC doesn’t have spurious local minima

Theorem 7.4 (Ge et al, 2016)

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

- all local min are global: for any local minimum \( \mathbf{U} \) of \( f(\mathbf{U}) \), it satisfies \( \mathbf{U} \mathbf{U}^\top = \mathbf{M} \);
- 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 $M \in \mathbb{R}^{n \times n}$ (square case for simplicity)
- $\text{rank}(M) = r \ll n$
- The thin Singular value decomposition (SVD) of $M$:

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

where $\Sigma = \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix}$ contain all singular values $\{\sigma_i\}$;

$U := [u_1, \ldots, u_r], \ V := [v_1, \ldots, v_r]$ consist of singular vectors
Signal + noise

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

One can write observation \(P_\Omega(M)\) as

\[
\frac{1}{p}P_\Omega(M) = \underbrace{M}_{\text{signal}} + \underbrace{\frac{1}{p}P_\Omega(M) - M}_{\text{noise}}
\]

- Noise has mean zero: \(\mathbb{E}\left[\frac{1}{p}P_\Omega(M)\right] = M\)
Low-rank denoising

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

**Algorithm 7.1 Spectral method**

\[ \hat{M} \leftarrow \text{best rank-}r \text{ approximation of} \ \frac{1}{p} \mathcal{P}_\Omega(M) \]

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

A $10^4 \times 10^4$ random rank-3 matrix $M$ 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

\[
\frac{\| \hat{M} - M \|_F}{\| M \|_F} \lesssim \frac{\max_{i,j} |M_{i,j}|}{\frac{1}{n} \| M \|_F} \cdot \sqrt{\frac{nr \log^2 n}{m}},
\]

\( \nu \) reflects whether energy of \( M \) is spread out, \( |M_{i,j}| \lesssim \mu r / n \);

• When \( m \gg \nu^2 n \log^2 n \), estimate \( \hat{M} \) is very close to truth\(^1\)

• Degrees of freedom \( \asymp nr \)
  \[ \rightarrow \text{nearly-optimal sample complexity for incoherent matrices} \]

\(^1\)The logarithmic factor can be improved.
To ensure $\hat{M}$ is a good estimate, it suffices to control noise $E$.

**Lemma 7.6**

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

$$
\|P_r(M + E) - M\| \leq 2\|E\|
$$

$$
\|P_r(M + E) - M\|_F \leq 2\sqrt{2r}\|E\|
$$

where $P_r(X)$ is the best rank-$r$ approximation of $X$. 
Lemma 7.7 (Weyl’s inequality, 1912)

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

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

Proof: Invoke the Courant-Fisher Minimax Characterization:

$$\sigma_k(A) = \max_{\dim(S) = k} \min_{0 \neq v \in S} \frac{\|Av\|_2}{\|v\|_2}.$$
Proof of Lemma 7.6

By matrix perturbation theory,

$$\left\| \mathcal{P}_r(M + E) - M \right\|$$

triangle inequality

$$\leq \left\| \mathcal{P}_r(M + E) - (M + E) \right\| + \left\| (M + E) - M \right\|$$

$$\leq \sigma_{r+1}(M + E) + \| E \|$$

Weyl's inequality

$$\leq \sigma_{r+1}(M) + \| E \| + \| E \| = 2\| E \|.$$
Controlling the noise

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

A bit of random matrix theory ...

**Lemma 7.8 (Chapter 2.3, Tao ’12)**

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

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

Then $\|X\| \lesssim \sqrt{n \log n}$. 
Proof of Theorem 7.5

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

\[
\text{Var}[\tilde{E}_{i,j}] = p(1 - p) \cdot \left( \frac{\sqrt{p}}{\mu/n} \cdot \frac{1}{p} M_{i,j} \right)^2 \leq \left( \frac{M_{i,j}}{\mu/n} \right)^2 \lesssim 1,
\]

\[
|\tilde{E}_{i,j}| \leq \frac{|M_{i,j}|}{\sqrt{p} \mu/n} \lesssim \frac{1}{\sqrt{p}},
\]

where we have used the fact

\[
|M_{i,j}| = |e_i^\top U \Sigma V^\top e_j| \leq \|U^\top e_i\| \cdot \sigma_1 \cdot \|V^\top e_j\| \quad \text{(by our assumptions)} \quad \lesssim \frac{\mu r}{n} \asymp \frac{\mu}{n}
\]

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

\[
\|\tilde{E}\| \lesssim \sqrt{n} \log n \quad \iff \quad \|E\| \lesssim \frac{\mu}{\sqrt{p} m} \log n
\]

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

\[
\text{minimize}_{U,V} \; f(U, V) := \|P_\Omega(UV^\top - M)\|_F^2.
\]

- **Gradient descent:** (our focus)
  \[
  U_{t+1} = P_U \left[U_t - \eta_t \nabla_U f(U_t, V_t)\right],
  \]
  \[
  V_{t+1} = P_V \left[V_t - \eta_t \nabla_V f(U_t, V_t)\right].
  \]

  where \(\eta_t\) is the step size and \(P_U, P_V\) denote the Euclidean projection onto some constraint sets;

- **Alternating minimization:** One optimizes \(U, V\) alternatively while fixing the other, which is a convex problem.
  \[
  U_{t+1} = \arg\min_U f(U, V_t),
  \]
  \[
  V_{t+1} = \arg\min_V f(U_{t+1}, V).
  \]
Gradient descent for matrix completion

\[
\text{minimize}_{X \in \mathbb{R}^{n \times r}} \quad f(X) = \sum_{(j,k) \in \Omega} \left( e_j^\top X X^\top e_k - M_{j,k} \right)^2
\]

**Algorithm 7.2 Gradient descent for MC**

**Input:** \( Y = [Y_{j,k}]_{1 \leq j, k \leq n}, r, p. \)

**Spectral initialization:** Let \( U^0 \Sigma^0 U^{0\top} \) be the rank-\( r \) eigendecomposition of

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

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

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

\[
X^{t+1} = X^t - \eta_t \nabla f \left( X^t \right).
\]
Gradient descent for matrix completion

Define the optimal transform from the $t$th iterate $X^t$ to $X^\natural$ as

$$Q^t := \arg\min_{R \in O_{r \times r}} \| X^t R - X^\natural \|_F.$$ 

**Theorem 7.9 (Ma et al., 2017)**

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

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

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

Gradient descent for matrix completion

Define the optimal transform from the $t$th iterate $X^t$ to $X^\dagger$ as

$$Q^t := \arg\min_{R \in O_{r \times r}} \|X^t R - X^\dagger\|_F.$$

**Theorem 7.9 (Ma et al., 2017)**

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

- $\|X^t Q^t - X^\dagger\|_F \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|X^\dagger\|_F$,  
- $\|X^t Q^t - X^\dagger\| \lesssim \rho^t \mu r \frac{1}{\sqrt{np}} \|X^\dagger\|$, \hspace{1cm} (spectral)  
- $\|X^t Q^t - X^\dagger\|_{2,\infty} \lesssim \rho^t \mu r \sqrt{\frac{\log n}{np}} \|X^\dagger\|_{2,\infty}$, \hspace{1cm} (incoherence)

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

- *Linear convergence* of $\|X^t X^{t\top} - M^\dagger\|$ in Frobenius, spectral and infinity norms.
Numerical evidence for noiseless data

Figure 7.1: Relative error of $X^t X^{t\top}$ (measured by $\|\cdot\|_F$, $\|\cdot\|$, $\|\cdot\|_\infty$) vs. iteration count for matrix completion, where $n = 1000$, $r = 10$, $p = 0.1$, and $\eta_t = 0.2$. 
Set \( \text{SNR} := \frac{\| M^b \|_F^2}{n^2 \sigma^2} \).

Figure 7.2: Squared relative error of the estimate \( \hat{X} \) (measured by \( \| \cdot \|_F, \| \cdot \|, \| \cdot \|_{2,\infty} \)) and \( \hat{M} = \hat{X} \hat{X}^\top \) (measured by \( \| \cdot \|_{\infty} \)) vs. SNR, where \( n = 500, r = 10, p = 0.1, \) and \( \eta_t = 0.2. \)
Lemma 7.10 (Restricted strong convexity and smoothness)

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

\[ \text{vec}(V)^\top \nabla^2 f(X) \text{vec}(V) \geq \frac{\sigma_{\min}}{2} \|V\|_F^2 \quad \text{(restricted strong convexity)} \]

\[ \|\nabla^2 f(X)\| \leq \frac{5}{2} \sigma_{\max} \quad \text{(smoothness)} \]

for all $X$ and $V = YH_Y - Z$, $H_Y := \arg\min_{R \in O_{r \times r}} \|YR - Z\|_F$ satisfying

- $\|X - X\|_{2,\infty} \leq \epsilon \|X\|_{2,\infty}$ \quad \text{(incoherence region)},
- $\|Z - X\| \leq \delta \|X\|,$

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

Restricted strong convexity and smoothness
Linear convergence induction I

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

$$\|X^{t+1}Q^{t+1} - X^\parallel\|_F \leq \|X^{t+1}Q^t - X^\parallel\|_F$$

(i) $$\|\left[ X^t - \eta \nabla f (X^t) \right] Q^t - X^\parallel\|_F$$

(ii) $$\|X^t Q^t - \eta \nabla f (X^t Q^t) - X^\parallel\|_F$$

(iii) $$\|X^t Q^t - \eta \nabla f (X^t Q^t) - \left( X^\parallel - \eta \nabla f (X^\parallel) \right) \|_F$$

where (i) follows from the GD rule, (ii) follows from the identity
\(\nabla f (X^t R) = \nabla f (X^t) \ R\) for any \(R \in \mathcal{O}^{r \times r}\), and (iii) follows from
\(\nabla f (X^\parallel) = 0\).
The fundamental theorem of calculus reveals

$$\begin{align*}
\text{vec} \left[ X^t Q^t - \eta \nabla f (X^t Q^t) - \left( X^\| - \eta \nabla f (X^\|) \right) \right] \\
= \text{vec} \left[ X^t Q^t - X^\| \right] - \eta \cdot \text{vec} \left[ \nabla f (X^t Q^t) - \nabla f (X^\|) \right] \\
= \left( I_{nr} - \eta \int_0^1 \nabla^2 f (X (\tau)) \, d\tau \right) \text{vec} \left( X^t Q^t - X^\| \right),
\end{align*}$$

(7.1)

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

$$\alpha^2 = \text{vec}(X^t Q^t - X^\|)^\top (I_{nr} - \eta A)^2 \text{vec}(X^t Q^t - X^\|)$$

$$\leq \left\| X^t Q^t - X^\| \right\|_F^2 + \eta^2 \| A \|^2 \left\| X^t Q^t - X^\| \right\|_F^2$$

$$- 2\eta \text{vec}(X^t Q^t - X^\|)^\top A \text{vec}(X^t Q^t - X^\|),$$

(7.2)
Based on the incoherence of $X^\perp$ and $X^t$, $\forall \tau \in [0, 1]$,

$$
\|X(\tau) - X^\perp\|_{2,\infty} \leq \|X^t Q^t - X^\perp\|_{2,\infty} \leq C \mu r \sqrt{\frac{\log n}{np}} \|X^\perp\|_{2,\infty}.
$$

\[\text{incoherence hypothesis}\]

Taking $X = X(\tau), Y = X^t$ and $Z = X^\perp$ 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,

$$
\text{vec}(X^t Q^t - X^\perp)^\top A \text{ vec}(X^t Q^t - X^\perp) \geq \frac{\sigma_{\min}}{2} \|X^t Q^t - X^\perp\|_F^2
$$

and

$$
\|A\| \leq \frac{5}{2} \sigma_{\max}.
$$
Substituting these two inequalities into (7.2) yields

\[
\alpha^2 \leq \left(1 + \frac{25}{4} \eta^2 \sigma_{\text{max}}^2 - \sigma_{\text{min}} \eta^2\right) \|X^t \hat{H}^t - X^\|$$_F^2
\]
\[
\leq \left(1 - \frac{\sigma_{\text{min}}}{2} \eta\right) \|X^t Q^t - X^\|$$_F^2
\]

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

\[
\alpha \leq \left(1 - \frac{\sigma_{\text{min}}}{4} \eta\right) \|X^t Q^t - X^\|$$_F.
\]

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


