# Matrix Completion: Theory and Implementations

Liangzu Peng

SIST, ShanghaiTech Univerisity

June 24, 2018

# What we did in this project

## Theory

A summary of Ge et al.[1].

[1] Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems (pp. 2973-2981).

[2] Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), 717.

[3] Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4), 1956-1982.

# What we did in this project

## Theory

A summary of Ge et al.[1].

## Implementations

[1] Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems (pp. 2973-2981).

[2] Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), 717.

[3] Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4), 1956-1982.

# What we did in this project

## Theory

A summary of Ge et al.[1].

## Implementations

- **PSD-GD** (as a sanity check for Ge et al., we implement **G**radient **D**ecent method for **PSD** matrix completion with some relaxations).

[1] Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems (pp. 2973-2981).

[2] Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), 717.

[3] Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4), 1956-1982.

# What we did in this project

## Theory

A summary of Ge et al.[1].

## Implementations

- **PSD-GD** (as a sanity check for Ge et al., we implement **G**radient **D**ecent method for **PSD** matrix completion with some relaxations).
- Nuclear Norm Regularized Minimization (Candes and Recht[2]).

[1] Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems (pp. 2973-2981).

[2] Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), 717.

[3] Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4), 1956-1982.

# What we did in this project

### Theory

A summary of Ge et al.[1].

### Implementations

- **PSD-GD** (as a sanity check for Ge et al., we implement **G**radient **D**ecent method for **PSD** matrix completion with some relaxations).
- Nuclear Norm Regularized Minimization (Candes and Recht[2]).
- **SVT** (Cai et al.[3]).

---

[1] Ge, R., Lee, J. D., & Ma, T. (2016). Matrix completion has no spurious local minimum. In Advances in Neural Information Processing Systems (pp. 2973-2981).

[2] Candès, E. J., & Recht, B. (2009). Exact matrix completion via convex optimization. Foundations of Computational mathematics, 9(6), 717.

[3] Cai, J. F., Candès, E. J., & Shen, Z. (2010). A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization, 20(4), 1956-1982.

# Theory

## Problem Setting (Ge et al.[1]).

$$\text{minimize}_X ||P_\Omega(XX^T - M)||_F^2,$$

where $M = ZZ^T$ is a positive semidefinite matrix, $\Omega$ is the set of observed entries and $P$ is the projection operator.

[4] Ongie, G., Willett, R., Nowak, R. D., & Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. arXiv preprint arXiv:1703.09631.

# Theory

Problem Setting (Ge et al.[1]).

$$\text{minimize}_X ||P_\Omega(XX^T - M)||_F^2,$$

where $M = ZZ^T$ is a positive semidefinite matrix, $\Omega$ is the set of observed entries and $P$ is the projection operator.

Solvability (Candes and Recht[2])

- Solvable for **low rank** matrix with **incoherence** assumption.

[4] Ongie, G., Willett, R., Nowak, R. D., & Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. arXiv preprint arXiv:1703.09631.

# Theory

## Problem Setting (Ge et al.[1]).

$$\text{minimize}_X ||P_\Omega(XX^T - M)||_F^2,$$

where $M = ZZ^T$ is a positive semidefinite matrix, $\Omega$ is the set of observed entries and $P$ is the projection operator.

## Solvability (Candes and Recht[2])

- Solvable for **low rank** matrix with **incoherence** assumption.
- Method for high rank matrix completion exists (algebraic geometry approach[4]).

[4] Ongie, G., Willett, R., Nowak, R. D., & Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. arXiv preprint arXiv:1703.09631.

# Theory

## Problem Setting (Ge et al.[1]).

$$\text{minimize}_X ||P_\Omega(XX^T - M)||_F^2,$$

where $M = ZZ^T$ is a positive semidefinite matrix, $\Omega$ is the set of observed entries and $P$ is the projection operator.

## Solvability (Candes and Recht[2])

- Solvable for **low rank** matrix with **incoherence** assumption.
- Method for high rank matrix completion exists (algebraic geometry approach[4]).
- Incoherence ball in Ge et al.[1] (rank-1 case): $\mathcal{B} = \{x : ||x||_\infty < \frac{2\mu}{\sqrt{d}}, ||x|| \leq 1\}$

[4] Ongie, G., Willett, R., Nowak, R. D., & Balzano, L. (2017). Algebraic variety models for high-rank matrix completion. arXiv preprint arXiv:1703.09631.

# Theory

Ge et al.[1] at first glance:

[5] Sun, R., & Luo, Z. Q. (2016). Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory, 62(11), 6535-6579.

# Theory

Ge et al.[1] at first glance:

- Goal: prove that **PSD** matrix completion, i.e., the following function

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 + \lambda R(X)$$

with $M = ZZ^T$ positive semidefinite **has no spurious local minimum** (i.e., local minimum=global minimum).

[5] Sun, R., & Luo, Z. Q. (2016). Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory, 62(11), 6535-6579.

# Theory

Ge et al.[1] at first glance:

- Goal: prove that **PSD** matrix completion, i.e., the following function

$$f(X) = \frac{1}{2}||P_{\Omega}(XX^T - M)||_F^2 + \lambda R(X)$$

  with $M = ZZ^T$ positive semidefinite **has no spurious local minimum** (i.e., local minimum=global minimum).
- Generalizable proof from a specific case to the general one, i.e.,
  - from incoherence ball $\mathcal{B}$ to $\mathbb{R}^{d \times d}$ (How?).
  - from rank-1 case to rank-$k$ case.

[5] Sun, R., & Luo, Z. Q. (2016). Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory, 62(11), 6535-6579.

# Theory

Ge et al.[1] at first glance:

- Goal: prove that **PSD** matrix completion, i.e., the following function

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 + \lambda R(X)$$

  with $M = ZZ^T$ positive semidefinite **has no spurious local minimum** (i.e., local minimum=global minimum).
- Generalizable proof from a specific case to the general one, i.e.,
  - from incoherence ball $\mathcal{B}$ to $\mathbb{R}^{d \times d}$ (How?).
  - from rank-1 case to rank-$k$ case.
- Use Lemma 3.1 in Sun and Luo[5],
  - by which it's enough to show $XX^T$ and $M = ZZ^T$ are close, e.g., in rank-$k$ case (informal),

$$||XX^T - ZZ^T||_F^2 \leq c, \text{ for some } c \in \mathbb{R}.$$

[5] Sun, R., & Luo, Z. Q. (2016). Guaranteed matrix completion via non-convex factorization. IEEE Transactions on Information Theory, 62(11), 6535-6579.

# Theory

**Highlights** in Ge et al.[1] (from $\mathcal{B}$ to $\mathbb{R}^{d \times d}$ (How?)):

# Theory

**Highlights** in Ge et al.[1] (from $\mathcal{B}$ to $\mathbb{R}^{d\times d}$ (How?)):

- Introduce a regularization term $R(X)$ to refine the geometry of the objective function, thus making every stationary point incoherent, e.g., in rank-1 case (informal),

$$R(x) = \Sigma_{i=1}^{d} h(x_i)$$
$$\overset{(*)}{\Rightarrow} ||x||_\infty \le c, \forall x \in \{x : \nabla f(x) + \nabla R(x) = 0\}$$

for some $c \in \mathbb{R}$, where $h(t) = (t-\alpha)^4 \mathbb{I}_{t\ge\alpha}$ for some $\alpha$ and (*) is Lemma 4.7 in Ge et al.[1].

# Theory

**Highlights** in Ge et al.[1] (from $\mathcal{B}$ to $\mathbb{R}^{d\times d}$ (How?)):

- Introduce a regularization term $R(X)$ to refine the geometry of the objective function, thus making every stationary point incoherent, e.g., in rank-1 case (informal),

$$R(x) = \Sigma_{i=1}^{d} h(x_i)$$
$$\overset{(*)}{\Rightarrow} ||x||_\infty \leq c, \forall x \in \{x : \nabla f(x) + \nabla R(x) = 0\}$$

  for some $c \in \mathbb{R}$, where $h(t) = (t-\alpha)^4 \mathbb{I}_{t\geq\alpha}$ for some $\alpha$ and (*) is Lemma 4.7 in Ge et al.[1].

- by which (which) if there is no spurious local minimum in the ball $\mathcal{B}$, a similar result can be obtained for the entire space $\mathbb{R}^{d\times d}$.

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?

A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?

A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.

**Note that**

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 \leq \frac{1}{2}||XX^T - P_\Omega(M)||_F^2 =: g(X).$$

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?
A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.
**Note that**

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 \leq \frac{1}{2}||XX^T - P_\Omega(M)||_F^2 =: g(X).$$

- It is easy to compute the gradient of $g(X)$ w.r.t. X. Indeed,

$$\nabla g(X) = 2XX^T X - 2P_\Omega(M)X.$$

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?
A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.
**Note that**

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 \leq \frac{1}{2}||XX^T - P_\Omega(M)||_F^2 =: g(X).$$

- It is easy to compute the gradient of $g(X)$ w.r.t. X. Indeed,

$$\nabla g(X) = 2XX^TX - 2P_\Omega(M)X.$$

- Minimizing $g(X)$ is very fast via gradient decent method.

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?
A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.
**Note that**

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 \leq \frac{1}{2}||XX^T - P_\Omega(M)||_F^2 =: g(X).$$

- It is easy to compute the gradient of $g(X)$ w.r.t. X. Indeed,

$$\nabla g(X) = 2XX^T X - 2P_\Omega(M)X.$$

- Minimizing $g(X)$ is very fast via gradient decent method.
- (assume that) Minimizing $g(X)$ as an upper bound of $f(X)$ will somehow minimize $f(X)$.

# Implementations

Q: I do not understand concentration inequality used in the proof, how do I believe the theory in Ge et al.[1]?
A: verify the theory empirically! $\Rightarrow$ **PSD-GD**.
**Note that**

$$f(X) = \frac{1}{2}||P_\Omega(XX^T - M)||_F^2 \leq \frac{1}{2}||XX^T - P_\Omega(M)||_F^2 =: g(X).$$

- It is easy to compute the gradient of $g(X)$ w.r.t. X. Indeed,

$$\nabla g(X) = 2XX^TX - 2P_\Omega(M)X.$$

- Minimizing $g(X)$ is very fast via gradient decent method.
- (assume that) Minimizing $g(X)$ as an upper bound of $f(X)$ will somehow minimize $f(X)$.
- Want to see **whether** $g(X)$ **always converges to the same point for random initialization**.

# Implementations

Experiments for **PSD-GD**.

- Use synthetic data (e.g., np.random).
- $M \in \mathbb{R}^{200\times 200}$.

# Implementations

Experiments for **PSD-GD**.

- Use synthetic data (e.g., np.random).
- $M \in \mathbb{R}^{200\times 200}$.

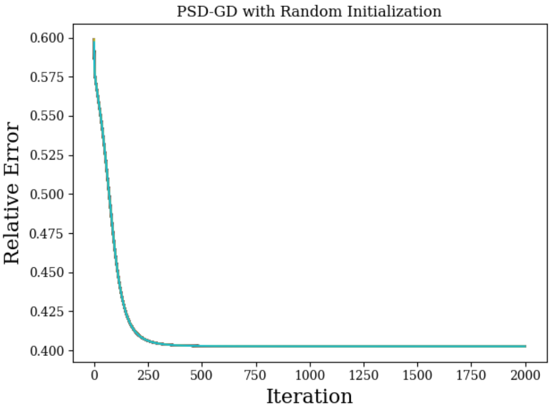


Figure 1: **PSD-GD** runs 500 times (500 curves in the plot). It can be seen that they all converge to the same function value.

# Implementations

Experiments for **PSD-GD**.

- Use synthetic data (e.g., np.random).
- $M \in \mathbb{R}^{200\times 200}$.

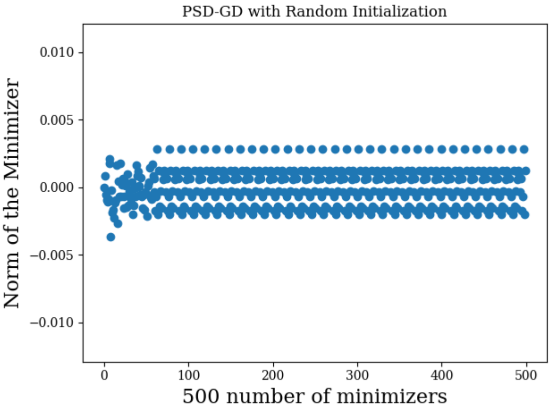


Figure 2: The norm of 500 convergent points. It can be seen that, up to some numerical errors ($\pm 0.005$), they have the same norm.

# Other Implementations

Nuclear Norm Minimization in Candes and Recht[2]:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & P_\Omega(X) = P_\Omega(M). \end{array}$$

# Other Implementations

Nuclear Norm Minimization in Candes and Recht[2]:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & P_\Omega(X) = P_\Omega(M). \end{array}$$

- Can be solved directly (e.g., SCS Solver in cvxpy).

# Other Implementations

Nuclear Norm Minimization in Candes and Recht[2]:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & P_\Omega(X) = P_\Omega(M). \end{array}$$

- Can be solved directly (e.g., SCS Solver in cvxpy).
- Can be (approximately) solved by Singular Value Thresholding Algorithm (Cai et al.[3]):

$$\begin{cases} X^k = \mathcal{D}_\tau(Y^{k-1}) \\ Y^k = Y^{k-1} + \lambda_k P_\Omega(M - X^k), \end{cases}$$

where $\mathcal{D}_\tau(Y) = prox_{\tau||\cdot||_*}(Y)$ (as shown in homework 4) and $\lambda_k \in (0, 2)$ is the stepsize.

# Other Implementations

Nuclear Norm Minimization in Candes and Recht[2]:

$$\begin{array}{ll} \text{minimize} & ||X||_* \\ \text{subject to} & P_\Omega(X) = P_\Omega(M). \end{array}$$

- Can be solved directly (e.g., SCS Solver in cvxpy).
- Can be (approximately) solved by Singular Value Thresholding Algorithm (Cai et al.[3]):

$$\begin{cases} X^k = \mathcal{D}_\tau(Y^{k-1}) \\ Y^k = Y^{k-1} + \lambda_k P_\Omega(M - X^k), \end{cases}$$

where $\mathcal{D}_\tau(Y) = prox_{\tau||\cdot||_*}(Y)$ (as shown in homework 4) and $\lambda_k \in (0,2)$ is the stepsize.

| Method | NucNorm | SVT |
|---|---|---|
| Time | 1437 s | 242 s |
| Relative Error | 0.0001 | 0.026 |

Table 1: Experiments running on MovieLens 100K ($n \times m = 943 \times 1682, |\Omega| = 10^5$), SVT trades accuracy for speed.

# Other Implementations

Experiments for **SVT** on MovieLens 100K dataset
($n \times m = 943 \times 1682, |\Omega| = 10^5$).

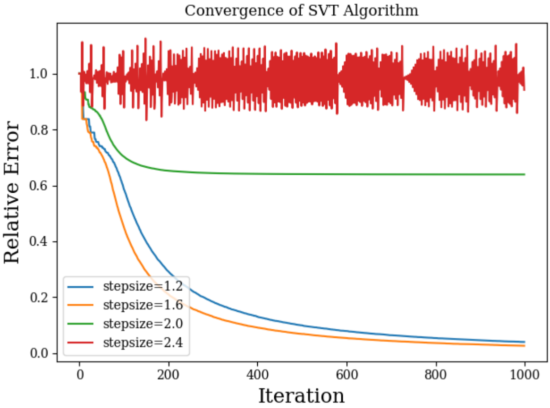


Figure 3: Corresponding to Theorem 4.2 in Cai et al.[3], **SVT** converges to a unique solution only when the stepsize $\lambda_k \in (0, 2)$.