A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Samples

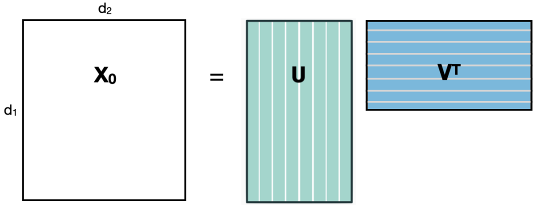
Christian Kümmerle
Johns Hopkins University

Claudio Mayrink Verdun
TU Munich

ICML 2021

Problem: Low-Rank Matrix Completion

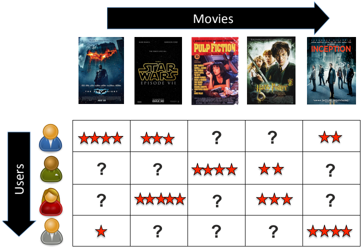
How to complete $d_1 d_2 - m$ missing entries of rank-r matrix $\mathbf{X}_0$



from a subset of m entries $y_\ell = (\mathbf{X}_0)_{i_\ell, j_\ell}$, with $\Omega = (i_\ell, j_\ell)_{\ell=1}^m$ index set of m locations?

Applications:

- Recommender systems



- Signal processing:
    - Sensor localization, . . .
- Dimensionality reduction

Algorithms for Low-Rank Matrix Completion

Since **(2003-)**: Large literature proposing algorithms for

$$\boxed{\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \operatorname{rank}(\mathbf{X}) \text{ s.t. } P_\Omega(\mathbf{X}) = y \quad \left(\text{with } P_\Omega : \mathbf{X} \mapsto (\mathbf{X}_{i_\ell,j_\ell})_{(i_\ell,j_\ell)\in\Omega}\right)}$$

\`\`Rank minimization´´: Challenging as objective non-convex and non-smooth!

Algorithms for Low-Rank Matrix Completion

Since **(2003-)**: Large literature proposing algorithms for

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \operatorname{rank}(\mathbf{X}) \text{ s.t. } P_\Omega(\mathbf{X}) = y \quad \left(\text{with } P_\Omega : \mathbf{X} \mapsto (\mathbf{X}_{i_\ell,j_\ell})_{(i_\ell,j_\ell)\in\Omega}\right)$$

''Rank minimization'': Challenging as objective non-convex and non-smooth!

Q: What should a good algorithm fulfill?

Algorithms for Low-Rank Matrix Completion

Since **(2003-)**: Large literature proposing algorithms for

$$\min_{\mathbf{X}\in\mathbb{R}^{d_1\times d_2}} \operatorname{rank}(\mathbf{X}) \text{ s.t. } P_\Omega(\mathbf{X}) = y \quad \left(\text{with } P_\Omega : \mathbf{X} \mapsto (\mathbf{X}_{i_\ell,j_\ell})_{(i_\ell,j_\ell)\in\Omega}\right)$$

\`\`Rank minimization'': Challenging as objective non-convex and non-smooth!

Q: What should a good algorithm fulfill?

- **Data-efficient:** Identify $\mathbf{X}_0$ from few samples, i.e., from m as small as possible, preferably from
$$m \approx \deg_{\mathbf{X}_0} = r(d_1 + d_2 - r).$$
- **Scalable:** Usable for large problems. Netflix prize data set: $d_1 \approx 480000$, $d_2 \approx 17000$ with $m \approx 10^8$.
- **Provable:** Guarantee solution of original problem under **realistic** assumptions.
- **Handle Ill-Conditioning:** $\kappa := \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0) \gg 1$.
Very common, e.g., in signal processing or discretization of PDEs.

Most Popular and Well-Studied Approaches

- **Convex optimization** (Nuclear norm minimization): $\min_{\mathbf{X}} \sum_i \sigma_i(\mathbf{X})$ s.t. $P_\Omega(\mathbf{X}) = y$.
  - **Data-efficiency:** $m > 3 \cdot \deg_{\mathbf{X}_0}$ necessary
  - **Scalability:**
  - **Guarantees:**
- Gradient Descent on **matrix factorization** (non-convex) **(Burer, Monteiro '03)**:
  - **Data-efficiency:**
  - **Scalability:**
  - **Guarantees:**
- **Riemannian optimization** **(Vandereycken '13, Boumal, Absil '15, Wei et al. '20)**:
  - **Data-efficiency:**
  - **Scalability:**
  - **Guarantees:**

Most Popular and Well-Studied Approaches

- **Convex optimization** (Nuclear norm minimization): $\min_{\mathbf{X}} \sum_i \sigma_i(\mathbf{X})$ s.t. $P_\Omega(\mathbf{X}) = y$.
  - **Data-efficiency:** $m > 3 \cdot \deg_{\mathbf{X}_0}$ necessary • **Scalability:** • **Guarantees:**
- Gradient Descent on **matrix factorization** (non-convex) **(Burer, Monteiro '03)**:
  - **Data-efficiency:** • **Scalability:** • **Guarantees:**
- **Riemannian optimization** **(Vandereycken '13, Boumal, Absil '15, Wei et al. '20)**:
  - **Data-efficiency:** • **Scalability:** • **Guarantees:**

Typical theoretical guarantees:

- Assume uniform random model for m sampling locations, μ_0-incoherent ground truth $\mathbf{X}_0 \in \mathbb{R}^{D \times D}$ of rank r, provide sufficient condition on m for convergence w.h.p.
- E.g., **(Chi, Liu, Li '20)** for GD on matrix fac.: $m = \Omega\left(\mu_0{}^2 \kappa^{14} r \deg_{\mathbf{X}_0} \log(D)\right)$, where condition number $\kappa := \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0)$. Thus, not applicable for $\kappa \gg 1$!

Are there any methods that complete very ill-conditioned low-rank matrices from few samples m?

Not really so far, but we propose a method (MatrixIRLS) to do this.

Our Approach: Non-Convex Rank Surrogates

Replace $\mathrm{rank}(X)$ by (smoothed) logdet-objective (as minimizers coincide very often):

$$\log\det(\mathbf{X}) = \sum_i \log(\sigma_i(\mathbf{X})) = \lim_{p\to 0} \sum_i \frac{\sigma_i(\mathbf{X})^p - 1}{p},$$

limit case of Schatten-p quasi-norm for $p \to 0$.

- **Prior work**: From concavity, smoothing + first order Taylor: Iteratively Reweighted Trace Minimization **(Fazel, Boyd, Hindi '03)** and Iteratively Reweighted Least Squares **(IRLS)** **(Fornasier, Rauhut, Ward '11)**, **(Mohan, Fazel '12)**
  - Data-efficiency: Methods are able to complete $\mathbf{X}_0$ from very few samples m.
  - Guarantees: Challenge: Non-convexity of F.
  - Scalability: Storage and SVDs of $O(d_1 d_2)$ matrices.

Our Approach: Matrix Iteratively Reweighted Least Squares

Our Contributions (**K. 19'**, **K**, **Mayrink Verdun '20**, **'21**):

- Propose IRLS method MatrixIRLS with weight operator that **utilities second-order/ curvature information** of smoothed rank surrogate (unlike the ones of **(Mohan, Fazel '12)**, **(Fornasier, Rauhut, Ward '11)**)
- Provide **guarantee:** Local convergence for minimal sample complexity $m = \Omega\left(\mu_0 \mathrm{deg}_{\mathbf{X}_0} \log(D)\right)$ with **locally quadratic** convergence rate ☺.

[1] N_{CG} : Nr. of inner iterations used in conjugate gradient solver of weighted least squares.

Our Approach: Matrix Iteratively Reweighted Least Squares

Our Contributions (**K. 19'**, **K**, **Mayrink Verdun '20**, **'21**):

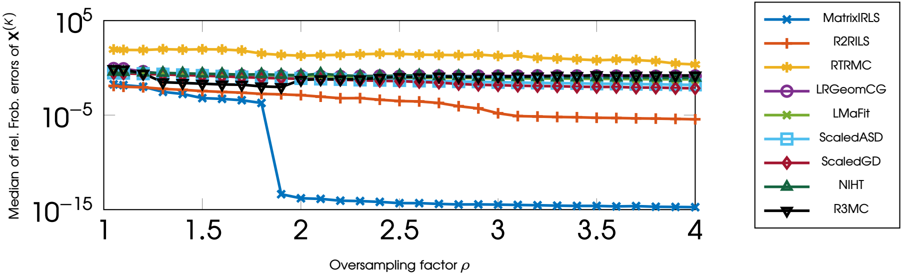
- Propose IRLS method MatrixIRLS with weight operator that **utilities second-order/ curvature information** of smoothed rank surrogate (unlike the ones of **(Mohan, Fazel '12)**, **(Fornasier, Rauhut, Ward '11)**)
- Provide **guarantee:** Local convergence for minimal sample complexity $m = \Omega\left(\mu_0 \mathrm{deg}_{\mathbf{x}_0} \log(D)\right)$ with **locally quadratic** convergence rate .
- Improve **scalability** by orders of magnitude compared to IRLS of **(Mohan, Fazel '12)**, **(Fornasier, Rauhut, Ward '11)** and **(K, Sigl '18)** :
  - Implicit representation of iterates in **low-rank + sparse format**, computed in time complexity $O\left((mr + r^2 D) \cdot N_{CG}\right)$, space complexity same as matrix factorization.[1]
  - Avoid ill-conditioning of weighted least-squares problems.

[1] N_{CG} : Nr. of inner iterations used in conjugate gradient solver of weighted least squares.

Empirical Performance: Very Ill-Conditioned Matrices

Experiment: Complete 1000×1000-matrix $\mathbf{X}_0$ of rank $r = 5$ with $\kappa := \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0) = 10^5$ from $m = \rho \cdot \deg_{\mathbf{X}_0}$ entries. **Observed:** Median relative Frobenius error of algorithm output over 100 realizations vs. oversampling factor ρ.

Comparison of MatrixIRLS to state-of-the-art algorithms R2RILS (**Bauch, Nadler, Zilber '21**), LRGeomCG (**Vandereycken '13**), RTRMC (**Boumal, Absil '15**), LMaFit (**Wen et al. '12**), ScaledASD (**Tanner, Wei '16**), ScaledGD (**Tong et al. '20**) and NIHT (**Tanner, Wei '13**), R3MC (**Mishra, Sepulchre '14**)



Summary

- Second-order methods for the optimization of non-convex rank surrogates rare in literature: We propose one such method, MatrixIRLS, attaining **state-of-the-art results** especially for low-rank matrix completion problems with small sample sizes that are ill-conditioned.
- IRLS (if done right) fits into a **sweet spot** for the optimization of very non-convex rank surrogates:
  Quadratic local convergence & fast escape from saddle points.
- Scalability of MatrixIRLS is comparable to (Burer-Monteiro type) matrix factorization approaches.

Summary

- Second-order methods for the optimization of non-convex rank surrogates rare in literature: We propose one such method, MatrixIRLS, attaining **state-of-the-art results** especially for low-rank matrix completion problems with small sample sizes that are ill-conditioned.
- IRLS (if done right) fits into a **sweet spot** for the optimization of very non-convex rank surrogates:
  Quadratic local convergence & fast escape from saddle points.
- Scalability of MatrixIRLS is comparable to (Burer-Monteiro type) matrix factorization approaches.

Caveat:

- Convergence guaranatee is only local, most guarantees for other algorithms are global.

Code available: `https://github.com/ckuemmerle/MatrixIRLS`.

Theoretical Guarantees for Matrix Completion Algorithms

Sufficient conditions on sample complexity m for uniform random sampling model, μ_0-incoherent ground truth $\mathbf{X}_0 \in \mathbb{R}^{D \times D}$ of rank r:

Nuclear Norm Min. **(Recht '11, Chen '15)**	$\Omega\left(\mu_0 \mathrm{deg}_{\mathbf{X}_0} \log^2(D)\right)$
OptSpace **(Keshavan, Montanari, Oh '10)**	$\Omega(\mu_0 \kappa^2 \mathrm{deg}_{\mathbf{X}_0} \max(\log(D), \kappa^4 r))$
AltMin **(Hardt, Wootters '15)**	$\Omega\left({\mu_0}^2 \log(\kappa) r^8 \mathrm{deg}_{\mathbf{X}_0} \log^2(D)\right)$
GD on matrix fac. **(Chi, Liu, Li '20)**	$\Omega\left({\mu_0}^2 \kappa^{14} r \mathrm{deg}_{\mathbf{X}_0} \log(D)\right)$
ScaledGD **(Tong, Ma, Chi '20)**	$\Omega\left(\mu_0 \kappa^2 r \mathrm{deg}_{\mathbf{X}_0} \max(\log(D), \mu_0 \kappa^2)\right)$
Necessary condition	$\Omega\left(\mu_0 \mathrm{deg}_{\mathbf{X}_0} \log(D)\right)$

Note: Large gap between necessary condition and guarantees for many methods if condition number $\kappa := \sigma_1(\mathbf{X}_0)/\sigma_r(\mathbf{X}_0) \gg 1$.

Main References

Christian Kümmerle, Claudio Mayrink Verdun
A Scalable Second Order Method for Ill-Conditioned Matrix Completion from Few Samples
ICML 2021.

Christian Kümmerle, Claudio Mayrink Verdun
Escaping Saddle Points in Ill-Conditioned Matrix Completion with a Scalable Second Order Method
Workshop on "Beyond first-order methods in ML systems", ICML 2020, arXiv:2009.02905.

Santiago Paternain, Aryan Mokhtari, Alejandro Ribeiro
A Newton-based method for nonconvex optimization with fast evasion of saddle points
SIAM Journal on Optimization, 29(1), 343-368, 2019

Karthik Mohan, Maryam Fazel
Iterative reweighted algorithms for matrix rank minimization
J. Mach. Learn. Res., vol. 13, pp. 3441--3473, 2012

Check out our code (including a collection of many MC algorithms) at:
`https://github.com/ckuemmerle/MatrixIRLS`.

Further References (I)

- Samuel Burer, Renato DC Monteiro
  A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization
  Mathematical Programming 95.2 (2003): 329-357.
- Bart Vandereycken
  Low-rank matrix completion by Riemannian optimization
  SIAM Journal on Optimization 23.2 (2013): 1214-1236.
- Nicholas Boumal, Pierre-Antoine Absil
  Low-rank matrix completion via preconditioned optimization on the Grassmann manifold
  Linear Algebra and its Applications 475.15 (2015): 200-239.
- Ke Wei, Jian-Feng Cai, Tony F. Chan, Shingyu Leung
  Guarantees of Riemannian optimization for low rank matrix completion
  Inverse Problems & Imaging, 14.2 (2020): 233-265.
- Massimo Fornasier and Holger Rauhut, Rachel Ward
  Low-rank Matrix Recovery via Iteratively Reweighted Least Squares Minimization
  SIAM Journal on Optimization, 21.4 (2011), 1614--1640.

Further References (II)

Maryam Fazel, Haitham Hindi, Stephen Boyd
Log-det heuristic for matrix rank minimization with applications to Hankel and Euclidean distance matrices
Proceedings of the American Control Conference, 3 (2003), 2156--2162.

Christian Kümmerle, Juliane Sigl
Harmonic Mean Iteratively Reweighted Least Squares for Low-Rank Matrix Recovery
Journal of Machine Learning Research, 19(47):1--49, 2018.

Christian Kümmerle
Understanding and Enhancing Data Recovery Algorithms: From Noise-Blind Sparse Recovery to Reweighted Methods for Low-Rank Matrix Optimization
Ph.D. Thesis, Technical University of Munich, 2019.

Ji Chen, Dekai Liu, Xiaodong Li
Nonconvex Rectangular Matrix Completion via Gradient Descent Without $\ell_{0,\infty}$-Regularization
IEEE Transactions on Information Theory 66.9 (2020): 5806-5841.

Tong, Tian, Cong Ma, Yuejie Chi
Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent
arXiv preprint arXiv:2005.08898 (2020).