Fine-Grained Analysis of Inductive Matrix Completion

Antoine Ledent[1], Rodrigo Alves[1], Yunwen Lei[2] and Marius Kloft[1]

[1]University of Kaiserslautern
[2]University of Birmingham

{ledent,alves,kloft}@cs.uni-kl.de y.lei@bham.ac.uk

October, 2021

Outline

Introduction and Problem Setting

Matrix Completion

Problem Setting: Unknown ground truth matrix $G \in \mathbb{R}^{m \times n}$.

Entries $G_{i,j}$ are observed $i.i.d.$ with (i, j) drawn from a sampling distribution $\mathcal{D}$. Can be observed with i.i.d. noise $\zeta \sim \mathcal{D}_n$.

Applications: Recommender Systems, drug interaction prediction, chemical engineering, social network analysis.

In Recommender Systems $G_{i,j}$ is the rating given by user i to item j.

Predictors: Functions F can be represented as the set of all their values on the entries $[m] \times [n]$: $F \in \mathbb{R}^{m \times n}$.

Loss: $l(F) = \mathbb{E}_{(i,j) \sim \mathcal{D}; \zeta \sim \mathcal{D}_n} l(F_{i,j}, R_{i,j} + \zeta)$.

Aim: Recover the ground truth G with high accuracy based on a small number of observations.

Low-rank Structure

Low-rank structure: In most applications it is reasonable to assume there is low-rank structure in the ground truth. In this case, it can be recovered with high accuracy from a small ($\ll mn$) number of observations.

Low-rank Structure

Low-rank structure: In most applications it is reasonable to assume there is low-rank structure in the ground truth. In this case, it can be recovered with high accuracy from a small ($\ll mn$) number of observations.

Explicit rank minimization: Srebro and Shraibman (2005) already shows that if the rank r is known, $\widetilde{O}\left(\frac{r(m+n)}{\epsilon^2}\right)$ entries are sufficient to recover the ground truth within ϵ accuracy.
However, explicitly minimizing the rank is NP hard.

Convex Relaxations

Convex relaxation (exact recovery): The nuclear norm $\|Z\|_*$ of a matrix Z (the sum of its singular values) indirectly promotes rank-sparsity. In Candès and Tao (2010), it was shown that $O(nr\log^2(n))$ entries are sufficient to recover R exactly with high probability via the following algorithm:

$$\begin{aligned} \min_{Z} \quad & \|Z\|_* \quad \text{subject to} \\ & Z_{i,j} = G_{i,j} \quad \forall (i,j) \in \Omega, \end{aligned} \tag{1}$$

where Ω is the set of observed entries and it is assumed that the entries are sampled uniformly at random.

Convex Relaxations

Convex relaxation (exact recovery): The nuclear norm $\|Z\|_*$ of a matrix Z (the sum of its singular values) indirectly promotes rank-sparsity. In Candès and Tao (2010), it was shown that $O(nr \log^2(n))$ entries are sufficient to recover R exactly with high probability via the following algorithm:

$$\begin{aligned} \min_{Z} \quad & \|Z\|_* \quad \text{subject to} \\ & Z_{i,j} = G_{i,j} \quad \forall (i,j) \in \Omega, \end{aligned} \tag{1}$$

where Ω is the set of observed entries and it is assumed that the entries are sampled uniformly at random.

Convex relaxation (noisy case): In practical scenarios, the nuclear norm can serve as a regularisor, as in the SoftImpute algorithm Mazumder et al. (2010):

$$\min_{Z \in \mathbb{R}^{m \times n}} \frac{1}{2} \|Z - G - \zeta\|_{\text{Fr}}^2 + \lambda \|Z\|_*. \tag{2}$$

Inductive Matrix Completion

Side information → feature vectors for the users (rows) and items (columns).

Collect in side information matrices $X \in \mathbb{R}^{m \times d_1}$ and $Y \in \mathbb{R}^{n \times d_2}$.

Optimization problem (Exact recovery):

$$\begin{aligned} \min_{M} \quad & \|M\|_* \quad \text{subject to} \\ \forall (i,j) \in \Omega, \quad & [XMY^\top]_{i,j} = G_{i,j}. \end{aligned} \tag{3}$$

Optimization problem (Approximate recovery):

$$\min_{M \in \mathbb{R}^{d_1 \times d_2}} \frac{1}{2} \|P_\Omega(XMY^\top - G - \zeta)\|_{\mathrm{Fr}}^2 + \lambda \|M\|_*. \tag{4}$$

State-of-the-art and Brief Summary of Our Contributions

Taxonomy of Theoretical Guarantees

In exact recovery, we assume the entries are observed exactly, and ask how many entries are required to recover the matrix exactly.

In approximate recovery, we use standard Rademacher analysis to prove generalisation bounds for a given loss function.

$\rightarrow$ Typically yields generalisation bounds of the order $O\left(\sqrt{\frac{f(r,d,m)}{N}}\right)$

$\rightarrow$ $O\left(\frac{f(r,d,m)}{\epsilon^2}\right)$ entries to reach ϵ expected loss.

Distributional assumptions: Can assume a uniform sampling or distribution-free setting.

Algorithm: SoftImpute, or modifications/ other regularisers. We focus here on slight modifications of problem (2), e.g. via weighting.

State-of-the art in MC

Table: Matrix completion results (trace norm-based only)

MC	Unif.Sampling	Distr.-free	Weighted version
Exact	$nr \log(n) \log(r)$	N/A	N/A
Approx.	$nr \log(n)$	$n^{3/2}\sqrt{r}$	$rn \log(n)$

(Cf. Candès and Tao (2010); Recht (2011); Candès and Recht (2009); Chen (2013); Foygel et al. (2011); Shamir and Shalev-Shwartz (2011))

State-of-the art in IMC

Table: Inductive matrix completion results (trace norm-based only)

IMC	Unif.Sampling	Distr.-free	Weighted
Exact	$rd\log(d)\log(n)$*	N/A	N/A
	$d^2r^3\log(d)$		
Appr. (sot)	rd^2	rd^2	None
Appr. (ours)	$rd\log(d)$	$d^{3/2}\sqrt{r}\log(d)$	$rd\log(d)$

(cf. Xu et al. (2013); Lu et al. (2016); Chiang et al. (2018); Jain and Dhillon (2013))
(* with orthogonal assumptions)

Our Contributions

1. We prove $O(rd\log(d))$ guarantees for approximate recovery IMC in the uniform sampling setting.

Our Contributions

1. We prove $O(rd\log(d))$ guarantees for approximate recovery IMC in the uniform sampling setting.
2. We prove $O(d^{3/2}\sqrt{r}\log(d))$ guarantees for approximate recovery IMC in the distribution-free sampling setting.

Our Contributions

1. We prove $O(rd\log(d))$ guarantees for approximate recovery IMC in the uniform sampling setting.
2. We prove $O(d^{3/2}\sqrt{r}\log(d))$ guarantees for approximate recovery IMC in the distribution-free sampling setting.
3. We introduce a weighted adaptation of the regulariser which brings the rate down to $O(rd\log(d))$, analogously to Foygel et al. (2011) (non inductive case).

Hints of Proof Techniques [Classic IMC]

Proof Strategy

$$\langle XMY^\top, R_N \rangle = \langle M, X^\top R_N Y \rangle. \tag{5}$$

Split by high and low variance entries of $O = X^\top R_N Y$.

Proof Strategy

$$\langle XMY^\top, R_N \rangle = \langle M, X^\top R_N Y \rangle. \tag{5}$$

Split by high and low variance entries of $O = X^\top R_N Y$.

Problem: Entries of $O = X^\top R_N Y$ are not independent! (cannot use standard concentration results).

Proof Strategy

$$\left\langle XMY^\top, R_N \right\rangle = \left\langle M, X^\top R_N Y \right\rangle. \tag{5}$$

Split by high and low variance entries of $O = X^\top R_N Y$.

Problem: Entries of $O = X^\top R_N Y$ are not independent! (cannot use standard concentration results).

Solution (technical): Iterative diagonalisation and peeling procedure on matrices $\mathbb{E}(\|O^\top O\|)$ and $\mathbb{E}(\|OO^\top\|)$ to use concentration results iteratively.

Final result (classic IMC)

W.p. $\geqslant 1 - \delta$ we have that $l(\hat{Z}) - \inf_{Z \in \mathcal{F}_\mathcal{M}} l(Z)$ is bounded by

$$\widetilde{O}\left[\sqrt{\frac{\ell b \mathbf{x}\mathbf{y}\mathcal{M}\sqrt{d}}{N}} + \frac{b}{\sqrt{N}} + \frac{\mathbf{x}\mathbf{y}\ell\mathcal{M} + \ell}{N}\right] + O\left(\sqrt{\frac{\log(1/\delta)}{N}}\right), \tag{6}$$

i.e.

$$\widetilde{O}\left[\max(b, \ell)\sqrt{\frac{\mathbf{x}\mathbf{y}\mathcal{M}\sqrt{d}}{N}}\right] \tag{7}$$

Fixing b, ℓ, $\mathcal{M} \sim \sqrt{d_1 d_1 r}$

$$\rightarrow \text{Rate of } \widetilde{O}\left(\frac{d^{3/2}\sqrt{r}}{\epsilon^2}\right)$$

Detailed Results and Proof Techniques [Weighted IMC]

Extending the Weighted Trace Norm to IMC

In Foygel et al. (2011) consider the marginal probabilities

$$p_i = \sum_j p_{i,j} \qquad q_j = \sum_i p_{i,j}.$$

$\hat{p}, \hat{q}$:empirical versions; $\tilde{p} = \alpha p + (1-\alpha)\frac{1}{m}$ (smoothed version); $\check{p}, \check{q}$: smoothed empirical versions.

Idea: regularise $\|F\|_{\tilde{p},\tilde{q}} := \|\sqrt{\tilde{p}}\sqrt{\tilde{q}}^\top \circ F\|_*$ or $\|F\|_{\check{p},\check{q}} := \|\sqrt{\check{p}}\sqrt{\check{q}}^\top \circ F\|_*$
$\rightarrow \widetilde{O}(rn)$ sample complexity.

Extending the Weighted Trace Norm to IMC

What about IMC?

Extending the Weighted Trace Norm to IMC

What about IMC?

In the general case, must again think about spectral structure of $X^\top R_N Y$, whose entries are not independent.

Need to consider interaction between the distribution $\mathcal{D}$ and and the privileged directions defined by the columns of X, Y.

Crucial quantities and first result

It turns out the eigenvalues of the following matrices play a key role in the generalization abilities of IMC:

$$X^\top \mathrm{diag}(q) X \qquad\qquad Y^\top \mathrm{diag}(\kappa) Y \tag{8}$$

$q, \kappa \sim$ marginal or empirical marginals.

Crucial quantities and first result

It turns out the eigenvalues of the following matrices play a key role in the generalization abilities of IMC:

$$X^\top \mathrm{diag}(q) X \qquad\qquad Y^\top \mathrm{diag}(\kappa) Y \tag{8}$$

$q, \kappa \sim$ marginal or empirical marginals.

We denote by σ^1, σ^2 the vectors containing the square roots of the eigenvalues of the matrices above, and σ^1_*, σ^2_* for the corresponding maxima.

We have proved that w.p. $\geqslant 1-\delta$, the generalisation gap $l(Z_S) - l(Z_*)$ is bounded by

$$\widetilde{O}\left(\frac{\ell}{\sqrt{N}}\mathcal{M}\max(\sigma^1_*, \sigma^2_*) + \frac{12\ell}{N}\mathcal{M}\mathbf{xy} + b\sqrt{\frac{\log(2/\delta)}{2N}}\right) \tag{9}$$

($\rightarrow$ sample complexity of $\widetilde{O}(rd)$ for uniform sampling)

Weighted Nuclear Norm for IMC

For the heavily non uniform case, we define diagonalize the matrices above as follows:

$$X^\top \mathrm{diag}(q) X = P^{-1} D P \qquad Y^\top \mathrm{diag}(\kappa) Y = Q^{-1} E Q \tag{10}$$

$$X^\top \mathrm{diag}(\hat{q}) X = \widehat{P}^{-1} \widehat{D} \widehat{P} \qquad Y^\top \mathrm{diag}(\hat{\kappa}) Y = \widehat{Q}^{-1} E \widehat{Q}. \tag{11}$$

We also apply a similar smoothing procedure: $\widetilde{D} = \frac{1}{2} D + \frac{1}{2d_1} I$, $\breve{D} = \frac{1}{2} \widehat{D} + \frac{1}{2d_1} I$ etc.

Weighted Nuclear Norm for IMC

For the heavily non uniform case, we define diagonalize the matrices above as follows:

$$X^\top \mathrm{diag}(q) X = P^{-1} D P \qquad Y^\top \mathrm{diag}(\kappa) Y = Q^{-1} E Q \tag{10}$$

$$X^\top \mathrm{diag}(\hat{q}) X = \widehat{P}^{-1} \widehat{D} \widehat{P} \qquad Y^\top \mathrm{diag}(\hat{\kappa}) Y = \widehat{Q}^{-1} E \widehat{Q}. \tag{11}$$

We also apply a similar smoothing procedure: $\widetilde{D} = \frac{1}{2} D + \frac{1}{2d_1} I$, $\breve{D} = \frac{1}{2} \widehat{D} + \frac{1}{2d_1} I$ etc.

We then propose to regularize the following norms (depending on whether the distribution is known)

$$\|\widetilde{M}\|_* := \|\widetilde{D}^{\frac{1}{2}} P M Q^{-1} \widetilde{E}^{\frac{1}{2}}\|_* \tag{12}$$

$$\|\breve{M}\|_* := \|\breve{D}^{\frac{1}{2}} \widehat{P} M \widehat{Q}^{-1} \breve{E}^{\frac{1}{2}}\|_*. \tag{13}$$

Summary of Results for Weighted IMC

We obtain:

- Sample complexity bounds of order $\widetilde{O}(rd)$ when assuming knowledge of the distribution.
- Sample complexity bounds of order $\widetilde{O}(rd)$ for the smoothed empirical setting (harder).

Practical Model and Experimental Results

Practical Model

In real life applications of IMC, it is very helpful to add a non inductive term, as originally proposed in Chiang et al. (2018).

Such modifications can be combined with our model:

$$\min_{M,Z} \frac{1}{N} \left\| P_\Omega \left(XMY^\top + Z - G - \zeta \right) \right\|_{\text{Fr}}^2 + \lambda_1 \| \check{D}^{\frac{1}{2}} \widehat{P} M \widehat{Q}^{-1} \check{E}^{\frac{1}{2}} \|_* + \lambda_2 \| \check{D}_I^{\frac{1}{2}} Z \check{E}_I^{\frac{1}{2}} \|_*$$

Experimental Results

Table: Results of real-world datasets (RMSE)

	S-I	**IMCNF**	**ATR-0.5**	**ATR-0.75**	**ATR-1.0**
Douban	0.9582	0.8197	0.7691	**0.7614**	0.8779
LastFM	2.4109	1.7612	**1.6159**	1.6943	2.3371
MovieLens	0.9280	0.9252	**0.9056**	0.9139	0.9262

S-I: **SoftImpute**, classic SoftImpute model Mazumder et al. (2010).
IMCNF: ((unweighted) model from Chiang et al. (2018)
ATR-α: our model with smoothing parameter α.

A wide range of synthetic data experiments are available in the paper.

Thank you

Neural networks and NTK

One of the main research directions we want to pursue can be summarized as follows:

Neural networks perform well even when the number of parameters is much larger than the number of samples, which is at odds with standard statistical learning theory. How can we explain this phenomenon?

The neural tangent kernel literature ((Jacot et al., 2018; Arora et al., 2019; Du et al., 2019) etc.) provides first (partial) answers:

Overparameterised networks trained with gradient descent behave like kernel methods as the number of parameters tends to infinity.

References I

S. Arora, S. Du, W. Hu, Z. Li, and R. Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, 2019.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9 (6):717, 2009.

E. J. Candès and T. Tao. The power of convex relaxation: Near-optimal matrix completion. *IEEE Trans. Inf. Theor.*, 56(5): 20532080, May 2010.

Y. Chen. Incoherence-optimal matrix completion. *Information Theory, IEEE Transactions on*, 61, 10 2013. doi: 10.1109/TIT.2015.2415195.

K.-Y. Chiang, I. S. Dhillon, and C.-J. Hsieh. Using side information to reliably learn low-rank matrices from missing and corrupted observations. *J. Mach. Learn. Res.*, 2018.

S. S. Du, K. Hou, R. R. Salakhutdinov, B. Poczos, R. Wang, and K. Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/663fd3c5144fd10bd5ca6611a9a5b92d-Paper.pdf.

R. Foygel, O. Shamir, N. Srebro, and R. R. Salakhutdinov. Learning with the weighted trace-norm under arbitrary sampling distributions. In J. Shawe-Taylor, R. S. Zemel, P. L. Bartlett, F. Pereira, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 24*, pages 2133–2141. Curran Associates, Inc., 2011. URL http://papers.nips.cc/paper/4303-learning-with-the-weighted-trace-norm-under-arbitrary-sampling-distributions.pdf.

A. Jacot, F. Gabriel, and C. Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/5a4be1fa34e62bb8a6ec6b91d2462f5a-Paper.pdf.

P. Jain and I. S. Dhillon. Provable inductive matrix completion. *CoRR*, abs/1306.0626, 2013.

J. Lu, G. Liang, J. Sun, and J. Bi. A sparse interactive model for matrix completion with side information. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper/2016/file/093b60fd0557804c8ba0cbf1453da22f-Paper.pdf.

References II

R. Mazumder, T. Hastie, and R. Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *J. Mach. Learn. Res.*, 11:22872322, Aug. 2010.

B. Recht. A simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12(null):34133430, Dec. 2011. ISSN 1532-4435.

O. Shamir and S. Shalev-Shwartz. Collaborative filtering with the trace norm: Learning, bounding, and transducing. In *Proceedings of the 24th Annual Conference on Learning Theory*, volume 19 of *Proceedings of Machine Learning Research*, pages 661–678. PMLR, 2011.

N. Srebro and A. Shraibman. Rank, trace-norm and max-norm. In P. Auer and R. Meir, editors, *Learning Theory*, pages 545–560, Berlin, Heidelberg, 2005. Springer Berlin Heidelberg. ISBN 978-3-540-31892-7.

M. Xu, R. Jin, and Z.-H. Zhou. Speedup matrix completion with side information: Application to multi-label learning. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS13, page 23012309, Red Hook, NY, USA, 2013. Curran Associates Inc.