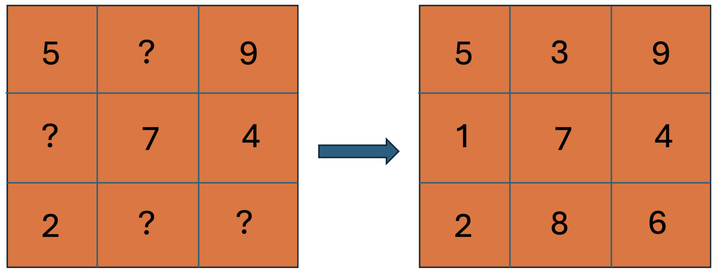
Alternating Minimization for Matrix Completion and Beyond

Yuzhou Gu*, Zhao Song†, Mingquan Ye‡, Junze Yin§, Lichen Zhang¶

∗ : NYU, † : Simons Institute for the Theory of Computing, ‡ : University of Illinois at Chicago, § : Rice University, ¶ : MIT CSAIL

Matrix Completion and Generalization

Matrix Completion. Given a matrix $M \in \mathbb{R}^{n\times n}$, suppose we are only able to observe $W \circ M$ where $\circ$ is the Hadamard product and $W \in \{0,1\}^{n\times n}$. The goal is to recover the matrix M through observation $W \circ M$.



In many practical scenarios, it is natural to assume M has low-rank (rank-k), so the goal is to compute a rank-k matrix $\widetilde{M}$ such that

$$\|\widetilde{M} - M\|_F \leq \epsilon$$

by observing $W \circ M$. Some additional assumptions have to be made in order for the problem to be approachable:

- Uniform sampling: we assume each entry of W follows an i.i.d. Bernoulli distribution with probability p;
- μ-incoherent: let $M = U\Sigma V^\top$ be its thin SVD, we assume $\max\{\|U_{i,*}\|_2^2, \|V_{i,*}\|_2^2\}_{i=1}^n \leq \mu \cdot \frac{k}{n}$.

Weighted Low-Rank Approximation. In matrix completion, we assume the ground truth M can be directly observed in terms of entries, but in practice, one usually can only observe a noisy version of M, captured by $M + N$ where N is a high-rank noise matrix.

In addition, when we know some entries are more important or some entries are noisy, we could use a tailored matrix W to facilitate the observation. Let $W \in \mathbb{R}^{n\times n}_{\geq 0}$, the weighted low-rank approximation problem asks one to observe $W \circ (M + N)$ and then compute a rank-k $\widetilde{M}$ such that

$$\|\widetilde{M} - M\|_F \leq \delta \cdot \|W \circ N\|_F + \epsilon.$$

Alternating Minimization

A particularly popular practical algorithm for this type of problem is *alternating minimization*, which could be succinctly described as follows:

- $U_0, V_0 \leftarrow \mathrm{SVD}(W \circ (M + N), k)$
- For $t = 1 \to T$
  - $U_t \leftarrow \arg\min_{U\in\mathbb{R}^{n\times k}} \|W \circ (M + N) - W \circ (UV_{t-1}^\top)\|_F^2$
  - $V_t \leftarrow \arg\min_{V\in\mathbb{R}^{n\times k}} \|W \circ (M + N) - W \circ (U_tV^\top)\|_F^2$
- Return $U_TV_T^\top$

It has many pros and cons:

- ✓: Commonly used in practice, easy to implement
- ✓: U_t, V_t can be computed approximately and efficiently
- ×: Requires to observe more entries than SDP-based method;
- ×: Theoretical analysis requires U_t, V_t to be computed *exactly*: it needs $O(|W| \cdot k^2 \log(1/\epsilon))$ time in theory.

Close the Gap

While practically efficient, the theory of alternating minimization is lacking. Our main goal is to close the gap between theory and practice:

- We want to design an alternating minimization algorithm whose updates could be efficiently approximated;
- We want to show the algorithm converges under approximate updates.

We achieve these two goals for both matrix completion and weighted low-rank approximation.

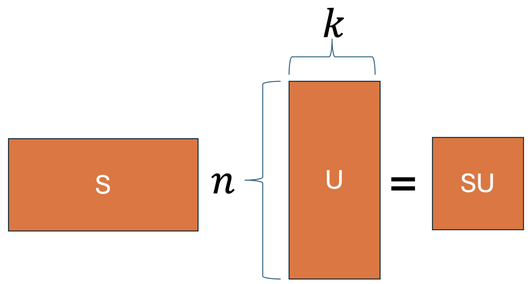
Theorem 1 (Gu, Song, Yin and Zhang, ICLR'24)**.** *Given a rank-k matrix $M \in \mathbb{R}^{n\times n}$ that is μ-incoherent, suppose $W \in \{0,1\}^{n\times n}$ with each entry being sampled with proper probability, there exists an alternating minimization algorithm that computes $\widetilde{M} \in \mathbb{R}^{n\times n}$ such that $\|\widetilde{M} - M\|_F \leq \epsilon$ for $\epsilon \in (0,1)$, in $\widetilde{O}(|W| \cdot k \log(1/\epsilon))$ time.*

Theorem 2 (Song, Ye, Yin and Zhang, ICLR'25)**.** *Given a rank-k matrix $M \in \mathbb{R}^{n\times n}$ that is μ-incoherent, a noise matrix $N \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{n\times n}_{\geq 0}$ satisfying mild conditions, there exists an alternating minimization algorithm that computes $\widetilde{M} \in \mathbb{R}^{n\times n}$ such that $\|\widetilde{M} - M\| \leq O(k\tau) \cdot \|W \circ N\| + \epsilon$ for $\epsilon \in (0,1)$ and τ is the condition number of M, in $\widetilde{O}(|W| \cdot k \log(1/\epsilon))$ time.*

In essence, we develop error robust framework for alternating minimization that could tolerate approximate updates, and we develop efficient algorithm to compute approximate updates in $\widetilde{O}(|W| \cdot k \log(1/\epsilon))$ time.

Algorithm: Sketch-and-Precondition

Sketch. To compute each update, note that if we let D_{W_i} to denote the diagonal matrix corresponds to the i-th column of W, compute each update is then solving n linear regressions in the form of $\min_{v\in\mathbb{R}^k} \|D_{W_i}(Uv - M_{*,i} - N_{*,i})\|_2^2$, and an efficient algorithmic approach is to apply a random *sketch* matrix: these matrices are structured random matrices that have much fewer rows than columns, can be applied efficiently, and preserve the cost of regression.



In particular, let $\mathrm{OPT} = \min_{v\in\mathbb{R}^k} \|Uv - b\|_2^2$, the matrix S satisfies

- $\min_{v\in\mathbb{R}^k} \|S(Uv - b)\|_2^2 \leq (1 + \epsilon) \cdot \mathrm{OPT}$;
- S has $\widetilde{O}(k/\epsilon^2)$ rows;
- S can be applied to U in $\widetilde{O}(nk + \mathrm{poly}(k, 1/\epsilon))$ time.

While it is tempting to draw a sketch matrix S and use it to solve the regression, approximately preserving the cost is not enough for our application. In fact, we want a vector $\widetilde{v}$ such that $\|\widetilde{v} - v_*\|_2$ is small where v_* is the optimal solution to the regression. This forces us to pick $\epsilon = 1/\mathrm{poly}(n, \tau)$, making the algorithm inefficient.

Algorithm: Sketch-and-Precondition

Precondition. To retain the efficiency of sketch-based approach for polynomially small ϵ, we instead use S as a *preconditioner*:

- Pick $\epsilon = 0.01$ for the sketch;
- Compute $SU = QR^{-1}$, the QR decomposition of SU;
- Use R as a preconditioner for the regression $\min_{v\in\mathbb{R}^k} \|Uv - b\|_2^2$;
- Compute an initial point v_0 by solving $\min_{v\in\mathbb{R}^k} \|SUv - Sb\|_2^2$, note that this gives a constant approximation;
- Use gradient descent to solve $\min_{v\in\mathbb{R}^k} \|URv - b\|_2^2$; with initial point v_0.

Since we start with a good initial point with a good preconditioner, the gradient descent converges in $\log(1/\epsilon)$ iterations. Moreover, each iteration could be implemented in $\widetilde{O}(nk)$ time. Put it together, we show how to solve one regression in $\widetilde{O}(nk \log(1/\epsilon))$ time. By further leveraging the sparsity pattern of W, we could improve the runtime for n regressions to $\widetilde{O}(|W| \cdot k \log(1/\epsilon))$.

Convergence: Perturb the Incoherence

To prove the algorithm converges, we first examine the approach for exact updates. Let U_*, V_* be the optimal rank-k factors for M, and let U_t, V_t be the exact updates for iteration t. The proof follows by

- Show that the initial distances $\mathrm{dist}(U_0, U_*), \mathrm{dist}(V_0, V_*)$ are bounded;
- Inductively:
  - Assume $\mathrm{dist}(V_{t-1}, V_*)$ is small and V_{t-1} is μ-incoherent;
  - Prove $\mathrm{dist}(U_t, U_*) \leq 1/4 \cdot \mathrm{dist}(U_{t-1}, U_*)$ and U_t is μ-incoherent;
  - Similarly, given $\mathrm{dist}(U_t, U_*)$ is small and U_t is μ-incoherent;
  - Prove $\mathrm{dist}(V_t, V_*) \leq 1/4 \cdot \mathrm{dist}(V_{t-1}, V_*)$ and V_t is μ-incoherent.

Since the distance shrinks by a constant factor at each iteration, it converges to ϵ-additive error in $\log(1/\epsilon)$ iterations. Since we instead compute approximate updates denoted by $\widetilde{U}_t, \widetilde{V}_t$ with the guarantees $\|\widetilde{U}_t - U_t\|, \|\widetilde{V}_t - V_t\|$ are small, we could use a similar approach to bound the distance (triangle inequality essentially). However, it is difficult to bound the incoherence of $\widetilde{U}_t, \widetilde{V}_t$, as the exact updates have closed-form solutions that are much easier to analyze. To circumvent this challenge, we develop a novel perturbation theory for incoherence: given $\|\widetilde{U}_t - U_t\|$ is small, we show the incoherence of $\widetilde{U}_t$ can be bounded by a factor of the incoherence of U_t. To do so, we note:

- The row norms of SVD factors of a matrix M could alternatively be rewritten as $\|(M^\top M)^{\dagger/2} M_{i,*}\|_2^2$, which are the statistical leverage scores of M;
- We could obtain a bound on the discrepancy between pseudo-inverses using known tools, and necessary ingredients are provided by the induction.

Since this also provides a perturbation theory for statistical leverage score, we hope it finds more applications.

Conclusion

We close the gap between theory and practice for alternating minimization, by providing nearly-linear time algorithms for matrix completion and weighted low-rank approximation, and a robust analytical framework to allow approximate updates in place of their exact counterpart. We also have experiments on moderate-sized matrices, and we show a 10%-20% speedup over the algorithm with exact updates.