LOW-RANK MATRIX RECOVERY

1. PROBLEM STATEMENT & MOTIVATION
2. SINGULAR VALUE DECOMPOSITION & LOW-RANK MATRIX APPROXIMATION
3. UNIQUENESS CONDITIONS
4. ALGORITHM & CONVERGENCE

---

1. PROBLEM STATEMENT & MOTIVATION

Suppose there exists $X^* \in \mathbb{R}^{m \times n}$, which is only observed as the vector $\vec{b} = \mathcal{A}(X_*)$, with $\mathcal{A} : \mathbb{R}^{m\times n} \to \mathbb{R}^d$, linear.

- Q: Can we recover X^* from $\vec{b}$ when $d \ll m, n$?

PROBLEM THAT DESERVES DISCUSSION, BECAUSE OF ITS PRACTICAL SIGNIFICANCE

e.g.: Netflix problem (1 MILLION $ PRIZE, 2009) Rating matrix, entry (i,j) rating of movie j by customer i, missing otherwise. Can we recover to make good recommendations?

- A: No. very ill-posed, underdetermined problem. $\mathcal{A}(X^*) = b_1$

We need other assumptions:

- $\boxed{\text{rank}(X^*) = r}$, low-rank (only few factors contribute to one's tastes) (Occam's Razor) $r \ll m, n$

(- uniform random sampling of observed entries) NO (Recht, Candès, Tao)

- number of observed entries

(- incoherence ...) NO

- Restricted Isometry Property

Existance: for general $\mathcal{A}$ and $\vec{b}$, the solution X^* might not exist. But it is a statistical problem, we think of recovering a matrix from $\vec{b}$

might be noised, you will not recover the same matrix

We suppose that such a matrix exists, $\vec{b}$ has been observed from it.

② SVD & LOW-RANK APPROXIMATION

Thm 1 Let $M \in \mathbb{R}^{m \times n}$, $r = \text{rank}(M)$. Then $M = U \Sigma V^T$, where $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n \times n}$ are orthogonal matrices, and

SVD: BELTRAMI 1873
JORDAN 1874
SYLVESTER 1889
proof: ECKART, YOUNG 1936

$$\Sigma = \left[\begin{array}{ccc|c} \sigma_1 & & 0 & \\ & \sigma_2 \ddots & & 0 \\ 0 & & \sigma_r & \\ \hline & 0 & & 0 \end{array}\right] \in \mathbb{R}^{m \times n}$$

where $\sigma_i \in \mathbb{R}$ and $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_r > 0$.

$$U = \underbrace{\begin{bmatrix} | & & | \\ u_1 & \cdots & u_m \\ | & & | \end{bmatrix}}_{\text{L.S.V.}} \qquad V = \underbrace{\begin{bmatrix} | & & | \\ v_1 & \cdots & v_n \\ | & & | \end{bmatrix}}_{\text{R.S.V.}}$$

v_i eig.vec. of $M^T M$
u_i eig.vec of $M M^T$

If $r = m$ or $r = n$, M is said to be full-rank.

Algorithm: Golub-Reinsch (1970), coût : $O(m n^2)$.

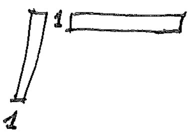
NB: IF M is square and σ_i are DISTINCT AND NONZERO, THEN U & V ARE UNIQUELY DETERMINED UP TO (COMPLEX) SIGNS.

PROPERTIES:

1. $$\|M\|_2 = \sup_{\substack{x \neq 0 \\ \in \mathbb{R}^n}} \frac{\|Mx\|_2}{\|x\|_2} = \sigma_1$$

2. $$\|M\|_F := \sqrt{\text{tr}(M^T M)} = \sqrt{\sum_i \sum_j M_{ij}^2} = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$$

3. "Rank-revealing" SVD : $$M = \sum_{i=1}^{r} \sigma_i \vec{u}_i \vec{v}_i^T$$



Storage is $(m+n+1)r$, instead of mn

Def. Low-rank matrix approximation is the optimization problem:

$$(P) \begin{cases} \min \|M - N\|_2 \\ \text{rank}(N) = k \ll \text{rank}(M) = r. \end{cases}$$

Thm 2 (ECKART-YOUNG (1936), SCHMIDT-MIRSKY). Solution N^* to (P) is given by the rank-k truncated SVD of M, namely

$$N^* = \sum_{i=1}^{k} \sigma_i(M) \vec{u}_i \vec{v}_i^T.$$

The minimal value is $\sigma_{k+1}(M)$. The minimizer N^* is unique $\iff$ $\sigma_k(M) > \sigma_{k+1}(M)$. We will use this thm later.
↑STRICT INEQ.

Let's go back to our problem.

- $\mathcal{A}$ is linear, therefore has a matrix representation. To see how, consider the ISOMORPHISM of MATRICES AS VECTORS, specified by the <u>vec operator</u>:

$$\mathrm{vec}: \mathbb{R}^{m\times n} \to \mathbb{R}^{mn} \quad \text{defined by}$$

$$X = [x_1, x_2 \dots x_n] \mapsto x = [x_1^T, x_2^T, \dots, x_n^T]^T$$

"COLUMNWISE STACKING of X"

Then $\exists A \in \mathbb{R}^{d\times mn}$ s.t. $\mathcal{A}(X) = A\,\mathrm{vec}(X)$.

(• For the Euclidean inner product: with $x = \mathrm{vec}(X)$, $y = \mathrm{vec}(Y)$

$$\langle x, y\rangle_{\mathbb{R}^{mn}} = x^T y = \mathrm{tr}(X^T Y) = \langle X, Y\rangle_{\mathbb{R}^{m\times n}}$$)

③ UNIQUENESS CONDITIONS

I) We need enough observations to recover X from $\vec{b}$.
(NECESSARY CONDITION)

Thm 3: If $d < (m+n-k)k$, then $\exists\, X \neq Y$ of rank $\leq k$ s.t. $\mathcal{A}(X) = \mathcal{A}(Y)$.

PROOF (?) NO, EXERCISE of LINEAR ALGEBRA: DIMENSION COUNTING & NULLITY + RANK

Consider the matrix subspace of $\mathbb{R}^{m\times n}$:

$$\mathcal{W} = \left\{ \begin{array}{c} \\ m \end{array}\!\!\begin{array}{cc} k & m-k \\ \left[U,\right. & \left. U_\perp\right] \end{array} \begin{array}{c} \\ k \\ m-k \end{array}\!\!\begin{array}{c} \begin{array}{cc} k & n-k \end{array} \\ \begin{bmatrix} C_1 & C_3 \\ C_2 & 0 \end{bmatrix} \end{array} \begin{array}{c} \\ n \end{array}\!\!\begin{array}{cc} k & n-k \\ \left[V,\right. & \left. V_\perp\right]^T \end{array}, \quad C_1, C_2, C_3 \text{ matrices of coefficients} \right\}$$

$d_w = \dim \mathcal{W} = (m+n-k)k \Rightarrow \exists\, W \in \mathbb{R}^{mn\times d_w}$, $\mathrm{rank}(W) = d_w$, whose columns span $\mathrm{vec}(Z)$, $Z \in \mathcal{W}$.

$$d\,\boxed{AW}^{\,d_w}$$

$\Rightarrow AW \in \mathbb{R}^{d\times d_w}$ has $\mathrm{Ker}(AW) \neq \{0\}$ if $d_w > d$ (NULLITY+RANK thm).

$\Rightarrow \exists\, \vec{c} \neq 0$, $\vec{c} \in \mathbb{R}^{d_w}$, s.t. $AW\vec{c} = 0$, $W\vec{c}$ is vectorized version of some element of $\mathcal{W}$.

$$\Leftrightarrow \mathcal{A}\left([U, U_\perp]\begin{bmatrix} C_1 & C_3 \\ C_2 & 0\end{bmatrix}[V, V_\perp]^T\right) = 0 \qquad (C_1, C_2, C_3 \text{ not all zeros})$$

$$\Leftrightarrow \mathcal{A}(UC_1V^T + UC_3V_\perp^T + U_\perp C_2 V^T) = 0$$

LINEARITY

$$\Leftrightarrow \mathcal{A}\Big(\underbrace{(UC_1 + U_\perp C_2)V^T}_{X,\ \mathrm{rank} \leq k}\Big) = \mathcal{A}\Big(\underbrace{-UC_3V_\perp^T}_{Y,\ \mathrm{rank}\leq k}\Big)$$

TWO MATRICES of rank $\leq k$ BELONGING TO TWO ORTHOG. SUBSPACES.

$X \neq Y$ ■

This is a <u>necessary</u> condition, but <u>not</u> <u>sufficient</u>.
$\Rightarrow$ look to $\mathcal{A}$ itself.

<u>Def.</u> $\mathcal{A}: \mathbb{R}^{m \times n} \to \mathbb{R}^d$, linear ($m \geq n$). $\mathcal{A}$ satisfies the rank-k RESTRICTED ISOMETRY PROPERTY (k-RIP) if $\exists$ smallest $0 < \delta_k < 1$ s.t.

(Candès, Tao, 2005)

$$(1-\delta_k)\|X\|_F^2 \leq \|\mathcal{A}(X)\|_2^2 \leq (1+\delta_k)\|X\|_F^2$$

for all matrices X of rank $\leq k$.

It looks a bit like s.v. characterization, but RESTRICTED to rank-k.

$$\sigma_n^2(A)\|x\|_2^2 \leq \|Ax\|_2^2 \leq \sigma_1^2(A)\|x\|_2^2 \quad \forall x$$

Heuristically: thinking of x as $\mathrm{vec}(X)$

$$\underbrace{\sigma_n^2(A)}_{(1-\delta_k)}\|X\|_F^2 \leq \|\overset{A\,\mathrm{vec}(X)}{\overset{\shortparallel}{\mathcal{A}(X)}}\|_2^2 \leq \underbrace{\sigma_1^2(A)}_{(1+\delta_k)}\|X\|_F^2 \quad \underbrace{\forall X}_{\mathrm{rank}(X) \leq k}$$

- Moreover ISOMETRIES ARE TRANSFORMATIONS THAT PRESERVE DISTANCES. RESTRICTED
- "W.R.T. rank-k MATRICES, $\mathcal{A}$ IS ACTING LIKE A NEAR-ISOMETRY!"

II) (SUFFICIENT CONDITION)

<u>Thm 4</u>: Suppose $\mathcal{A}$ is $2k$-RIP with $\delta_{2k} < 1$. Then X_* is the <u>ONLY</u> matrix of rank $\leq k$ s.t. $b = \mathcal{A}(X_*)$.

<u>Proof</u>: by contradiction, assume $\exists\, Y \neq X_*$ s.t. $\mathcal{A}(Y) = b$ with $\mathrm{rank}(Y) \leq k$. Then by linearity of $\mathcal{A}$:

$$\mathcal{A}(\underline{X_* - Y}) = 0$$

rank $\leq 2k$ (because of "rank subadditivity")

So we can use $2k$-RIP ("$\mathcal{A}$ is $2k$-RIP by assumption")

$$0 = \|\mathcal{A}(X_* - Y)\|_2^2 \geq \underbrace{(1-\delta_{2k})}_{>0}\underbrace{\|X_* - Y\|_F^2}_{>0} > 0 \quad \lightning \quad \blacksquare$$

"So $\mathcal{A}$ 2K-RIP, ⇒ we can recover X_* from $\vec{b}$"

But HOW? What is the algorithm?

④ ITERATIVE HARD THRESHOLDING

Idea: minimize misfit of X w.r.t. $\vec{b}$ using the obj. fun.

$$f(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$$

We could minimize $f(X)$ with the steepest descent (easiest optimization method)

$i \geq 0$

X_0 given

$$X_{i+1} = X_i \underbrace{- \alpha_i \nabla f(X_i)}_{\text{DIRECTION of STEEPEST DESCENT}}$$

α_i: stepsize, chosen to have sufficient decrease (e.g. using exact line-search, Armijo backtracking)

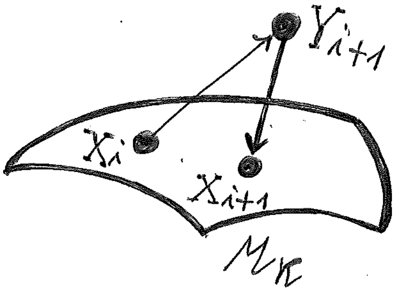
Converges to a stationary point of f: $\nabla f(X_i) \to 0$ $i \to \infty$.

Observe: "if we start with X_0 of rank k, rank$(X_1) > k$ in general!" Eventually → full-rank matrix!

⇒ NO HOPE IN FINDING X_* as X_∞!

How do we fix it? FORCE X_i to be rank k:

$$X_{i+1} = \Pi_k(X_i - \alpha_i \nabla f(X_i))$$



Π_k: projection onto the set of matrices of rank at most k M_k

$$\Pi_k(Y) = \underset{\text{rank}(Z) \leq k}{\text{argmin}} \|Z - Y\|$$

$\overset{(\text{Thm } 2)}{=}$ best rank-k approximation of Y

$\overset{(\text{Thm } 2)}{=}$ truncated svd of Y

PROJECTED S.D. = I.H.T.

Convergence?

Lemma. Let $\mathcal{A}$ be $2K$-RIP. Then IHT with $X_0 = 0$ and $\alpha_i = \alpha = 1/(1+\delta_{2K})$ satisfies:

$$f(X_{i+1}) \leq f(X_*) + \frac{\delta_{2K}}{1-\delta_{2K}} \|\mathcal{A}(X_* - X_i)\|_2^2 ,$$

with $\operatorname{rank}(X_*) \leq K$.

Proof (?) uses Taylor series of $f(X_{i+1})$, $2K$-RIP of $\mathcal{A}$, and best rank approx.

Thm 5 (Convergence to global optimal X^*).
Let $\vec{b} = \mathcal{A}(X_*)$ for some rank-K **matrix** X_*. Under assumptions of LEMMA, IHT satisfies

$$\|\mathcal{A}(X_{i+1}) - b\|_2^2 \leq \rho_K \|\mathcal{A}(X_i) - b\|_2^2$$

with $\rho_K = \dfrac{2\delta_{2K}}{1-\delta_{2K}} < 1$ if $\delta_{2K} < 1/3$.

Proof: use Lemma, and $\vec{b} = \mathcal{A}(X_*)$:

$$f(X_{i+1}) \leq \underbrace{f(X_*)}_{=0} + \frac{\delta_{2K}}{1-\delta_{2K}} \underbrace{\|\underbrace{\mathcal{A}(X_* - X_i)}_{b - \mathcal{A}(X_i)}\|_2^2}_{2f(X_i)}$$

$$f(X_{i+1}) \leq \frac{2\delta_{2K}}{1-\delta_{2K}} f(X_i)$$

$\rho_K < 1$ if $\delta_{2K} < 1/3$

$$\Rightarrow f(X_{i+1}) \leq \rho_K^{i+1} f(X_0)$$

so the obj. function $f(X_i) \to 0$ as $i \to \infty$, hence $X_i \to X_*$ since the global minimiser is unique by thm. 4. ∎

RATE of CONVERGENCE of **IHT** is exponential (geometric), soberly called linear convergence in optimization.

This shows the convergence of the method.