On Momentum Acceleration for Randomized Coordinate Descent in Matrix Completion

Matthew Callahan[1], Trung Vu[2], and Raviv Raich[1]

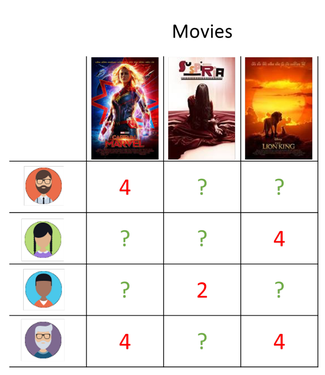
[1] School of EECS, Oregon State University, Corvallis, OR 97331-5501, USA
[2] Department of CSEE, University of Maryland, Baltimore County, MD 21250-0002, USA
callamat@orgonstate.edu

Matrix Completion

Goal: given *i.* Ω, a set of known entries of an $m \times n$ matrix $\boldsymbol{X}$ and *ii.* its rank r, fill-in the remaining entries:

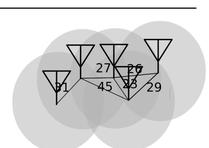
$$\min_{\boldsymbol{X}\in\mathbb{R}^{m\times n}} \frac{1}{2}\|\mathcal{P}_\Omega\left(\boldsymbol{X}-\boldsymbol{M}\right)\|_F^2$$
subject to: $\text{rank}(\boldsymbol{X}) \le r$

$$[\mathcal{P}_\Omega(\boldsymbol{X})]_{ij} = \begin{cases} X_{ij} & (i,j)\in\Omega, \\ 0 & (i,j)\notin\Omega. \end{cases}$$



Applications

- Recommender Systems [1]
- Image Inpainting [2]
- Device Localization [3]
- **Challenge:** The matrix can be high-dimensional



Node	1	2	3	4	5
1	0	31	?	?	?
2	31	0	27	45	?
3	?	27	0	23	26
4	?	45	23	0	29
5	?	?	26	29	0

Randomized Coordinate Descent (RCD) for Matrix Completion

- Unconstrained reformulation:

$$\min_{\boldsymbol{A}\in\mathbb{R}^{m\times r},\boldsymbol{B}\in\mathbb{R}^{n\times r}} \frac{1}{2}\|\mathcal{P}_\Omega\left(\boldsymbol{A}\boldsymbol{B}^T-\boldsymbol{M}\right)\|_F^2$$

- Algorithm [11] $\mathcal{O}(|\Omega|r)$:
  1. Minimize with respect to a coordinate of $\boldsymbol{A}^{(k)}$ or of $\boldsymbol{B}^{(k)}$
  2. After $(n+m)r$ repetitions of 1., refactor $\boldsymbol{A}^{(k)}\boldsymbol{B}^{(k)^T} = \boldsymbol{X}^{(k)}$

- Refactor as
$$\boldsymbol{A}^{(k)} = \boldsymbol{U}^{(k)}\sqrt{\boldsymbol{\Sigma}^{(k)}}, \qquad \boldsymbol{B}^{(k)} = \boldsymbol{V}^{(k)}\sqrt{\boldsymbol{\Sigma}^{(k)}}$$
where the SVD of $\boldsymbol{X}^{(k)}$ is $\boldsymbol{U}^{(k)}\boldsymbol{\Sigma}^{(k)}\boldsymbol{V}^{(k)^T}$

RCD Error Analysis [11]

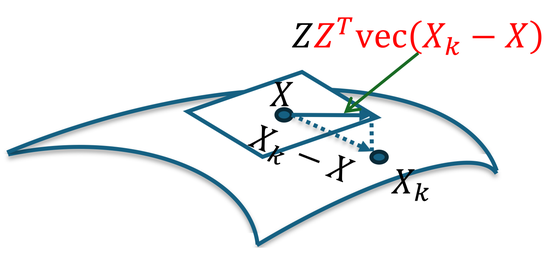
The expected value of the projected error $\boldsymbol{\delta}^{(k)} = (\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{Z})^{1/2}\boldsymbol{Z}^T(\text{vec}(\boldsymbol{A}^{(k)}\boldsymbol{B}^{(k)^T}-\boldsymbol{M}))$ follows:

$$\mathbb{E}[\boldsymbol{\delta}^{(k+1)}] = \underbrace{\mathbb{E}\left[\boldsymbol{I}-\frac{\boldsymbol{q}^*(\boldsymbol{q}^*)^T}{(\boldsymbol{q}^*)^T\boldsymbol{q}^*}\right]}_{\boldsymbol{T}}\mathbb{E}[\boldsymbol{\delta}^{(k)}] + o(\|\boldsymbol{\delta}^{(k)}\|^2)$$

$$\boldsymbol{q}^* = \begin{cases} \sqrt{\sigma_j}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{Z})^{1/2}\boldsymbol{Z}^T(\boldsymbol{v}_j\otimes\boldsymbol{e}_i^{(m)}) & \text{for } \boldsymbol{A}_{ij}^{(k)}, \\ \sqrt{\sigma_j}(\boldsymbol{Z}^T\boldsymbol{S}\boldsymbol{S}^T\boldsymbol{Z})^{1/2}\boldsymbol{Z}^T(\boldsymbol{e}_i^{(n)}\otimes\boldsymbol{u}_j) & \text{for } \boldsymbol{B}_{ij}^{(k)} \end{cases}$$

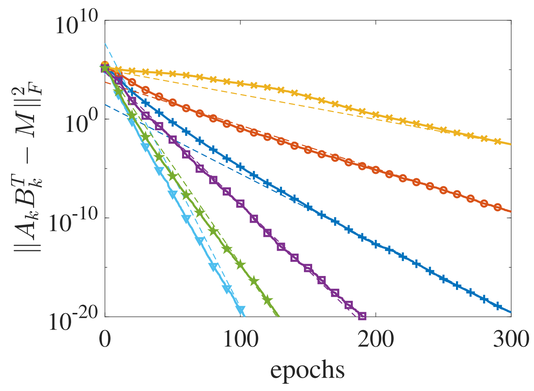
Key result:
Linear convergence rate $\rho(\boldsymbol{T})$ when $\rho(\boldsymbol{T}) < 1$

Goal: Develop efficient tuning-free accelerated RCD



Rank r manifold

RCD Error Analysis Experiments



Comparison of asymptotic rate (dashed) to empirical rate (solid) on 120×100 matrices

Polyak's Momentum Acceleration

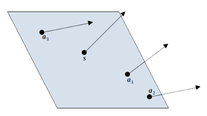
Fixed point iteration:
$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k), \quad \text{for} \quad k = 0, 1, 2, \ldots,$$
Error:
$$\boldsymbol{\epsilon}_{k+1} = \boldsymbol{T}\boldsymbol{\epsilon}_k + \boldsymbol{q}(\boldsymbol{\epsilon}_k), \quad \|\boldsymbol{q}(\boldsymbol{\epsilon})\| \le q\|\boldsymbol{\epsilon}\|^2$$
Convergence rate:
$$\rho = \rho(\boldsymbol{T})$$
Accelerated fixed point:
$$\boldsymbol{x}_{k+1} = \boldsymbol{f}(\boldsymbol{x}_k) + \beta(\boldsymbol{x}_k - \boldsymbol{x}_{k-1}).$$
Accelerated error:
$$\begin{bmatrix}\boldsymbol{\epsilon}_{k+1}\\ \boldsymbol{\epsilon}_k\end{bmatrix} = \boldsymbol{H}(\beta)\begin{bmatrix}\boldsymbol{\epsilon}_k\\ \boldsymbol{\epsilon}_{k-1}\end{bmatrix} + \begin{bmatrix}\boldsymbol{q}(\boldsymbol{\epsilon}_k)\\ \boldsymbol{q}(\boldsymbol{\epsilon}_{k-1})\end{bmatrix}$$
$$\boldsymbol{H}(\beta) = \begin{bmatrix}\boldsymbol{T}+\beta\boldsymbol{I} & -\beta\boldsymbol{I}\\ \boldsymbol{I} & \boldsymbol{0}\end{bmatrix}$$
Convergence rate with optimal β^*:
$$\rho = 1 - \sqrt{1-\rho(\boldsymbol{T})}$$

Momentum Accelerated RCD

1. Initialize $\boldsymbol{A}_0$, $\boldsymbol{B}_0$
2. Set $\boldsymbol{A}_1 = \boldsymbol{A}_0$, $\boldsymbol{B}_1 = \boldsymbol{B}_0$
3. for $k = 1, 2, \ldots$
   1. Generate $(\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{B}}_k)$ from iterated RCD:
   $$(\tilde{\boldsymbol{A}}_k, \tilde{\boldsymbol{B}}_k) = \underbrace{\text{RCD}(\text{RCD}(\ldots\text{RCD}(\boldsymbol{A}_k, \boldsymbol{B}_k)))}_{t \text{ nested functions}} \quad \mathcal{O}(t|\Omega|r)$$
   2. Update the factors as follows:
   $$\boldsymbol{A}_{k+1} = \tilde{\boldsymbol{A}}_k + \beta(\boldsymbol{A}_k - \boldsymbol{A}_{k-1}) \quad \mathcal{O}(mr),$$
   $$\boldsymbol{B}_{k+1} = \tilde{\boldsymbol{B}}_k + \beta(\boldsymbol{B}_k - \boldsymbol{B}_{k-1}) \quad \mathcal{O}(nr).$$

Modified Refactorization

- RCD applies refactorization for unique solution
$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \text{Refactor}(\boldsymbol{A}, \boldsymbol{B})$$



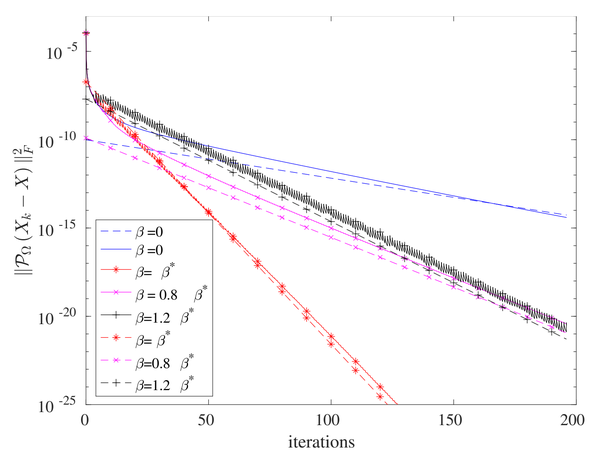
- Requirement: Maintain sign consistency
- Solution: Pick direction vector $\boldsymbol{s}$ and ensure alignment of A on same side of $\boldsymbol{s}$ hyperplane

$$\tilde{\boldsymbol{A}} = \hat{\boldsymbol{A}}\text{diag}(\text{sign}(\hat{A}^T\boldsymbol{s}))$$
$$\tilde{\boldsymbol{B}} = \hat{\boldsymbol{B}}\text{diag}(\text{sign}(\hat{A}^T\boldsymbol{s}))$$

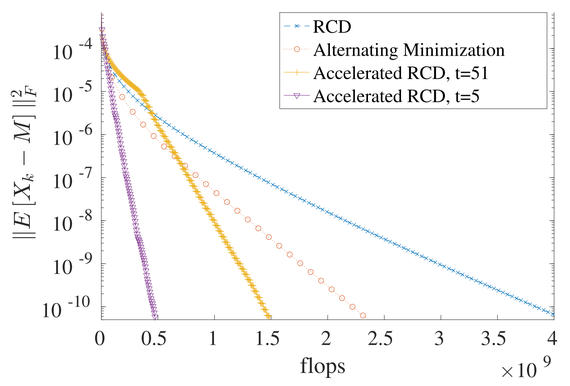
Analysis and Stepsize Selection

- Unaccelerated t-epoch convergence rate:
$$\rho_t = \lambda_{max}(\boldsymbol{I}-\boldsymbol{Q})^{(n+m)rt}$$
- Optimal momentum selection:
$$\beta^* = (1-\sqrt{1-\rho_t})^2$$
- Improved convergence results:
$$\|\mathbb{E}[\boldsymbol{A}_k\boldsymbol{B}_k^T - \boldsymbol{M}]\|_F \le C(1-\sqrt{1-\rho_t})^k$$
- Per-epoch computation complexity remains
$$\mathcal{O}(|\Omega|r)$$

Results



Twice objective value vs. epochs for accelerated randomized coordinate descent with various values of momentum parameter. Number of epochs per acceleration step is $t = 51$.



Squared Frobenius error of mean distance for an 80×80 matrix vs. flops for different algorithms. The algorithms shown are the unaccelerated RCD ($*$), optimally accelerated RCD ($+$ and ∇), and alternating minimization ($\circ$).

Conclusion

- Polyak's momentum acceleration to RCD provides an efficient method
- Epoch-level acceleration preserves RCD complexity
- Tight analysis of unaccelerated RCD allows for optimal hyper-parameter selection