Divide-and-Conquer Matrix Factorization

Lester Mackey[†]

Collaborators:

Ameet Talwalkar[‡] Michael I. Jordan[††]

[†]Stanford University [‡]UCLA [††]UC Berkeley

December 14, 2015

Motivation: Large-scale Matrix Completion

Goal: Estimate a matrix $\mathbf{L}_0 \in \mathbb{R}^{m\times n}$ given a subset of its entries

$$\begin{bmatrix} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 & 1 & \dots & 4 \\ 3 & 4 & 5 & \dots & 1 \\ 2 & 5 & 3 & \dots & 5 \end{bmatrix}$$

Examples

- Collaborative filtering: How will user i rate movie j?
  - Netflix: 40 million users, 200K movies and television shows
- Ranking on the web: Is URL j relevant to user i?
  - Google News: millions of articles, 1 billion users
- Link prediction: Is user i friends with user j?
  - Facebook: 1.5 billion users

Motivation: Large-scale Matrix Completion

Goal: Estimate a matrix $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ given a subset of its entries

$$\begin{bmatrix} ? & ? & 1 & \dots & 4 \\ 3 & ? & ? & \dots & ? \\ ? & 5 & ? & \dots & 5 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 3 & 1 & \dots & 4 \\ 3 & 4 & 5 & \dots & 1 \\ 2 & 5 & 3 & \dots & 5 \end{bmatrix}$$

State of the art MC algorithms

- Strong estimation guarantees
- Plagued by expensive subroutines (e.g., truncated SVD)

This talk

- Present divide and conquer approaches for scaling up any MC algorithm while maintaining strong estimation guarantees

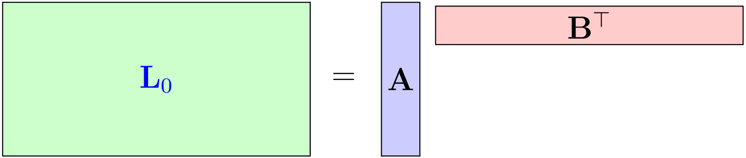
Exact Matrix Completion

Goal: Estimate a matrix $\mathbf{L}_0 \in \mathbb{R}^{m \times n}$ given a subset of its entries

Noisy Matrix Completion

Goal: Given entries from a matrix $\mathbf{M} = \mathbf{L}_0 + \mathbf{Z} \in \mathbb{R}^{m \times n}$ where $\mathbf{Z}$ is entrywise noise **and** $\mathbf{L}_0$ **has rank** $\mathbf{r} \ll m, n$, estimate $\mathbf{L}_0$

- Good news: $\mathbf{L}_0$ has $\sim (m+n)r \ll mn$ degrees of freedom



  - Factored form: $\mathbf{A}\mathbf{B}^\top$ for $\mathbf{A} \in \mathbb{R}^{m \times r}$ and $\mathbf{B} \in \mathbb{R}^{n \times r}$
- Bad news: Not all low-rank matrices can be recovered

Question: What can go wrong?

What can go wrong?

Entire column missing

$$\begin{bmatrix} 1 & 2 & ? & 3 & \ldots & 4 \\ 3 & 5 & ? & 4 & \ldots & 1 \\ 2 & 5 & ? & 2 & \ldots & 5 \end{bmatrix}$$

- No hope of recovery!

Solution: Uniform observation model

Assume that the set of s observed entries Ω is drawn uniformly at random:

$$\Omega \sim \mathrm{Unif}(m, n, s)$$

What can go wrong?

Bad spread of information

$$\mathbf{L} = \begin{bmatrix}1\\0\\0\end{bmatrix}\begin{bmatrix}1\end{bmatrix}\begin{bmatrix}1 & 0 & 0\end{bmatrix} = \begin{bmatrix}1 & 0 & 0\\0 & 0 & 0\\0 & 0 & 0\end{bmatrix}$$

- Can only recover $\mathbf{L}$ if $\mathbf{L}_{11}$ is observed

Solution: Incoherence with standard basis (Candès and Recht, 2009)

A matrix $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top \in \mathbb{R}^{m\times n}$ with $\mathrm{rank}(\mathbf{L}) = r$ is *incoherent* if

Singular vectors are not too skewed: $\begin{cases}\max_i \|\mathbf{U}\mathbf{U}^\top\mathbf{e}_i\|^2 \le \mu r/m\\ \max_i \|\mathbf{V}\mathbf{V}^\top\mathbf{e}_i\|^2 \le \mu r/n\end{cases}$

and not too cross-correlated: $\|\mathbf{U}\mathbf{V}^\top\|_\infty \le \sqrt{\frac{\mu r}{mn}}$

(In this literature, it's good to be incoherent)

How do we estimate $\mathbf{L}_0$?

First attempt:

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \operatorname{rank}(\mathbf{A}) \\ \text{subject to} & \sum_{(i,j)\in\Omega}(\mathbf{A}_{ij} - \mathbf{M}_{ij})^2 \leq \Delta^2. \end{array}$$

Problem: Computationally intractable!

Solution: Solve **convex** relaxation (Fazel, Hindi, and Boyd, 2001; Candès and Plan, 2010)

$$\begin{array}{ll} \text{minimize}_{\mathbf{A}} & \|\mathbf{A}\|_* \\ \text{subject to} & \sum_{(i,j)\in\Omega}(\mathbf{A}_{ij} - \mathbf{M}_{ij})^2 \leq \Delta^2 \end{array}$$

where $\|\mathbf{A}\|_* = \sum_k \sigma_k(\mathbf{A})$ is the trace/nuclear norm of $\mathbf{A}$.

Questions:

- Will the nuclear norm heuristic successfully recover $\mathbf{L}_0$?
- Can nuclear norm minimization scale to large MC problems?

Noisy Nuclear Norm Heuristic: Does it work?

Yes, with high probability.

Typical Theorem

If $\mathbf{L}_0$ with rank r is incoherent, $s \gtrsim rn\log^2(n)$ entries of $\mathbf{M} \in \mathbb{R}^{m\times n}$ are observed uniformly at random, and $\hat{\mathbf{L}}$ solves the noisy nuclear norm heuristic, then

$$\|\hat{\mathbf{L}} - \mathbf{L}_0\|_F \leq f(m,n)\Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$.

- See Candès and Plan (2010); Mackey, Talwalkar, and Jordan (2014b); Keshavan, Montanari, and Oh (2010); Negahban and Wainwright (2010)
- Implies **exact** recovery in the noiseless setting ($\Delta = 0$)

Noisy Nuclear Norm Heuristic: Does it scale?

Not quite...

- Standard interior point methods (Candès and Recht, 2009):
  $O(|\Omega|(m+n)^3 + |\Omega|^2(m+n)^2 + |\Omega|^3)$
- More efficient, tailored algorithms:
  - Singular Value Thresholding (SVT) (Cai, Candès, and Shen, 2010)
  - Augmented Lagrange Multiplier (ALM) (Lin, Chen, Wu, and Ma, 2009a)
  - Accelerated Proximal Gradient (APG) (Toh and Yun, 2010)
  - All require rank-k truncated SVD on **every** iteration

Take away: These provably accurate MC algorithms are too expensive for large-scale or real-time matrix completion

Question: How can we scale up a given matrix completion algorithm and still retain estimation guarantees?

Divide-Factor-Combine (DFC)

Our Solution: Divide and conquer

1. Divide $\mathbf{M}$ into submatrices.
2. Factor each submatrix **in parallel**.
3. Combine submatrix estimates to estimate $\mathbf{L}_0$.

Advantages

- Submatrix completion is often much cheaper than completing $\mathbf{M}$
- Multiple submatrix completions can be carried out in parallel
- DFC works with **any** base MC algorithm
- With the right choice of division and recombination, yields estimation guarantees comparable to those of the base algorithm

DFC-Proj: Partition and Project

1. Randomly partition $\mathbf{M}$ into t column submatrices $\mathbf{M} = \begin{bmatrix}\mathbf{C}_1 & \mathbf{C}_2 & \cdots & \mathbf{C}_t\end{bmatrix}$ where each $\mathbf{C}_i \in \mathbb{R}^{m\times l}$
2. Complete the submatrices **in parallel** to obtain
$$\begin{bmatrix}\hat{\mathbf{C}}_1 & \hat{\mathbf{C}}_2 & \cdots & \hat{\mathbf{C}}_t\end{bmatrix}$$
   - Reduced cost: Expect t-fold speed-up per iteration
   - Parallel computation: Pay cost of one cheaper MC
3. Project submatrices onto a single low-dimensional column space
   - Estimate column space of $\mathbf{L}_0$ with column space of $\hat{\mathbf{C}}_1$
$$\hat{\mathbf{L}}^{proj} = \hat{\mathbf{C}}_1\hat{\mathbf{C}}_1^{+}\begin{bmatrix}\hat{\mathbf{C}}_1 & \hat{\mathbf{C}}_2 & \cdots & \hat{\mathbf{C}}_t\end{bmatrix}$$
   - Common technique for randomized low-rank approximation (Frieze, Kannan, and Vempala, 1998)
   - Minimal cost: $O(mk^2 + lk^2)$ where $k = \mathrm{rank}(\hat{\mathbf{L}}^{proj})$
4. **Ensemble:** Project onto column space of each $\hat{\mathbf{C}}_j$ and average

DFC: Does it work?

Yes, with high probability.

Theorem (Mackey, Talwalkar, and Jordan, 2014b)

If $\mathbf{L}_0$ with rank r is incoherent and $s = \omega(r^2 n \log^2(n)/\epsilon^2)$ entries of $\mathbf{M} \in \mathbb{R}^{m\times n}$ are observed uniformly at random, then $l = o(n)$ random columns suffice to have

$$\|\hat{\mathbf{L}}^{proj} - \mathbf{L}_0\|_F \leq (2+\epsilon) f(m,n)\Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0\|_F \leq \Delta$ and the noisy nuclear norm heuristic is used as a base algorithm.

- Can sample vanishingly small fraction of columns $(l/n \to 0)$
- Implies exact recovery for noiseless $(\Delta = 0)$ setting
- Analysis streamlined by matrix Bernstein inequality

DFC: Does it work?

Yes, with high probability.

Proof Ideas:

1. If $\mathbf{L}_0$ is incoherent (has good spread of information), its partitioned submatrices are incoherent w.h.p.
2. Each submatrix has sufficiently many observed entries w.h.p.

$\Rightarrow$ Submatrix completion succeeds

3. Random submatrix captures the full column space of $\mathbf{L}_0$ w.h.p.
   - Analysis builds on randomized ℓ_2 regression work of Drineas, Mahoney, and Muthukrishnan (2008)

$\Rightarrow$ Column projection succeeds

DFC Noisy Recovery Error

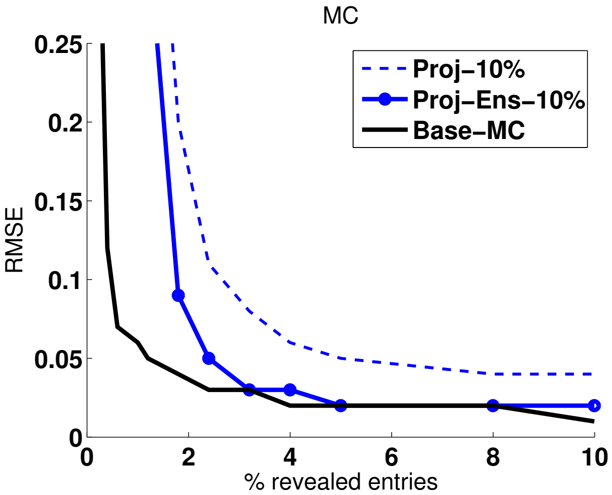


Figure : Recovery error of DFC relative to base algorithm (APG) with $m = 10K$ and $r = 10$.

DFC Speed-up

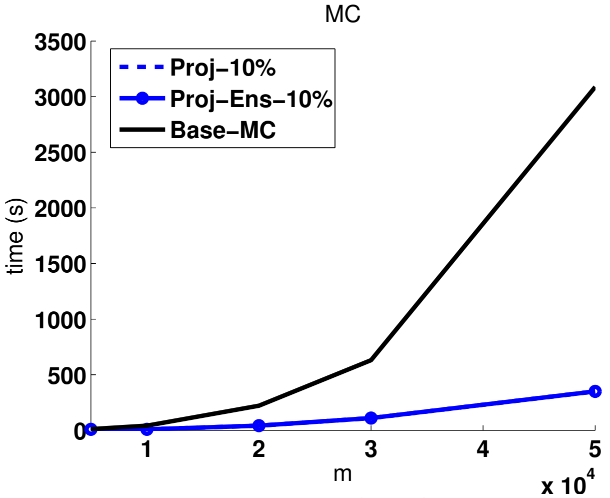


Figure : Speed-up over base algorithm (APG) for random matrices with $r = 0.001m$ and 4% of entries revealed.

Application: Collaborative filtering

Task: Given a sparsely observed matrix of user-item ratings, predict the unobserved ratings

Issues

- Full-rank rating matrix
- Noisy, non-uniform observations

The Data

- **Netflix Prize Dataset**[1]
    - 100 million ratings in $\{1, \ldots, 5\}$
    - 17,770 movies, 480,189 users

[1] http://www.netflixprize.com/

Application: Collaborative filtering

Task: Predict unobserved user-item ratings

Method	**Netflix**	
	RMSE	**Time**
APG	0.8433	2653.1s
DFC-Proj-25%	0.8436	689.5s
DFC-Proj-10%	0.8484	289.7s
DFC-Proj-Ens-25%	0.8411	689.5s
DFC-Proj-Ens-10%	0.8433	289.7s

Robust Matrix Factorization

Goal: Given a matrix $\mathbf{M} = \mathbf{L}_0 + \mathbf{S}_0 + \mathbf{Z}$ where $\mathbf{L}_0$ is low-rank, $\mathbf{S}_0$ is sparse, and $\mathbf{Z}$ is entrywise noise, recover $\mathbf{L}_0$ (Chandrasekaran, Sanghavi, Parrilo, and Willsky, 2009; Candès, Li, Ma, and Wright, 2011; Zhou, Li, Wright, Candès, and Ma, 2010)

Examples:

- Background modeling/foreground activity detection

$\mathbf{M}$ $\mathbf{L}_0$ $\mathbf{S}$

(Candès, Li, Ma, and Wright, 2011)

Robust Matrix Factorization

Goal: Given a matrix $\mathbf{M} = \mathbf{L}_0 + \mathbf{S}_0 + \mathbf{Z}$ where $\mathbf{L}_0$ is low-rank, $\mathbf{S}_0$ is sparse, and $\mathbf{Z}$ is entrywise noise, recover $\mathbf{L}_0$ (Chandrasekaran, Sanghavi, Parrilo, and Willsky, 2009; Candès, Li, Ma, and Wright, 2011; Zhou, Li, Wright, Candès, and Ma, 2010)

$\mathbf{M}$

$\mathbf{L}_0$

$\mathbf{S}_0$

- $\mathbf{S}_0$ can be viewed as an outlier/gross corruption matrix
  - Ordinary PCA breaks down in this setting
- **Harder than MC:** outlier locations are unknown
- **More expensive than MC:** dense, fully observed matrices

How do we recover $\mathbf{L}_0$?

First attempt:

$$\begin{aligned} \text{minimize}_{\mathbf{L},\mathbf{S}} \quad & \operatorname{rank}(\mathbf{L}) + \lambda \operatorname{card}(\mathbf{S}) \\ \text{subject to} \quad & \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \le \Delta. \end{aligned}$$

Problem: Computationally intractable!

Solution: Convex relaxation

$$\begin{aligned} \text{minimize}_{\mathbf{L},\mathbf{S}} \quad & \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \\ \text{subject to} \quad & \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \le \Delta. \end{aligned}$$

where $\|\mathbf{S}\|_1 = \sum_{ij} \mathbf{S}_{ij}$ is the ℓ_1 entrywise norm of $\mathbf{S}$.

Question: Does it work?

- Will noisy *Principal Component Pursuit* (PCP) recover $\mathbf{L}_0$?

Question: Is it efficient?

- Can noisy PCP scale to large RMF problems?

Noisy Principal Component Pursuit: Does it work?

Yes, with high probability.

Theorem (Zhou, Li, Wright, Candès, and Ma, 2010)

If $\mathbf{L}_0$ with rank r is incoherent, and $\mathbf{S}_0 \in \mathbb{R}^{m \times n}$ contains s non-zero entries with uniformly distributed locations, then if

$$r = O\left(m / \log^2 n\right) \quad \text{and} \quad s \leq c \cdot mn,$$

the minimizer to the problem

$$\begin{aligned} \text{minimize}_{\mathbf{L},\mathbf{S}} \quad & \|\mathbf{L}\|_* + \lambda \|\mathbf{S}\|_1 \\ \text{subject to} \quad & \|\mathbf{M} - \mathbf{L} - \mathbf{S}\|_F \leq \Delta. \end{aligned}$$

with $\lambda = 1/\sqrt{n}$ satisfies

$$\|\hat{\mathbf{L}} - \mathbf{L}_0\|_F \leq f(m,n)\Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0 - \mathbf{S}_0\|_F \leq \Delta$.

- See also Agarwal, Negahban, and Wainwright (2011)

Noisy Principal Component Pursuit: Is it efficient?

Not quite...

- Standard interior point methods: $O(n^6)$ (Chandrasekaran, Sanghavi, Parrilo, and Willsky, 2009)
- More efficient, tailored algorithms:
  - Accelerated Proximal Gradient (APG) (Lin, Ganesh, Wright, Wu, Chen, and Ma, 2009b)
  - Augmented Lagrange Multiplier (ALM) (Lin, Chen, Wu, and Ma, 2009a)
  - Require rank-k truncated SVD on **every** iteration
  - Best case $\mathrm{SVD}(m, n, k) = O(mnk)$

Idea: Leverage the divide-and-conquer techniques developed for MC in the RMF setting

DFC: Does it work?

Yes, with high probability.

Theorem (Mackey, Talwalkar, and Jordan, 2014b)

If $\mathbf{L}_0$ with rank r is incoherent, and $\mathbf{S}_0 \in \mathbb{R}^{m \times n}$ contains $s \leq c \cdot mn$ non-zero entries with uniformly distributed locations, then

$$l = O\left(\frac{r^2 \log^2(n)}{\epsilon^2}\right)$$

random columns suffice to have

$$\|\hat{\mathbf{L}}^{proj} - \mathbf{L}_0\|_F \leq (2+\epsilon) f(m,n) \Delta$$

with high probability when $\|\mathbf{M} - \mathbf{L}_0 - \mathbf{S}_0\|_F \leq \Delta$ and noisy principal component pursuit is used as the base algorithm.

- Can sample polylogarithmic number of columns
- Implies exact recovery for noiseless ($\Delta = 0$) setting

DFC Estimation Error

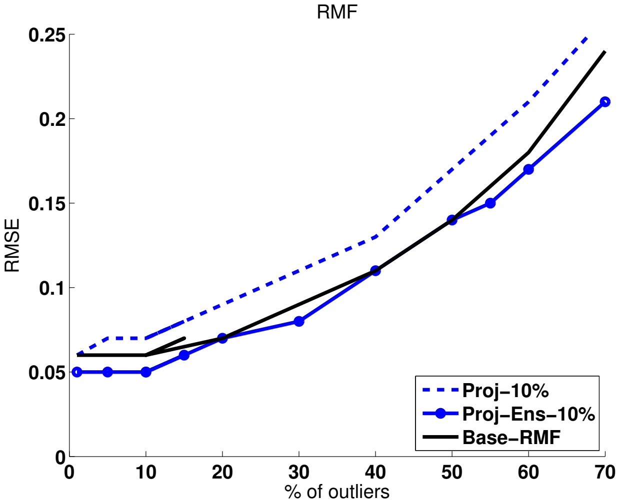


Figure : Estimation error of DFC and base algorithm (APG) with $m = 1K$ and $r = 10$.

DFC Speed-up

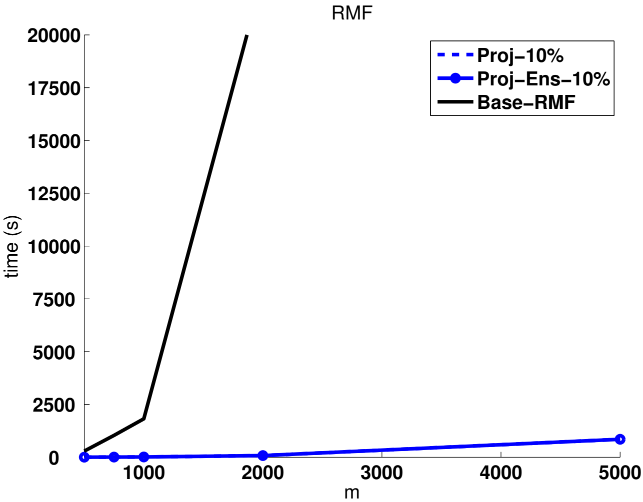


Figure : Speed-up over base algorithm (APG) for random matrices with $r = 0.01m$ and 10% of entries corrupted.

Application: Video background modeling

Task

- Each video frame forms one column of matrix $\mathbf{M}$
- Decompose $\mathbf{M}$ into stationary background $\mathbf{L}_0$ and moving foreground objects $\mathbf{S}_0$

$\mathbf{M}$

$\mathbf{L}_0$

$\mathbf{S}_0$

Challenges

- Video is noisy
- Foreground corruption is often clustered, not uniform

Application: Video background modeling

Example: Significant foreground variation

Specs

- 1 minute of airport surveillance (Li, Huang, Gu, and Tian, 2004)
- 1000 frames, 25344 pixels
- Base algorithm: half an hour
- DFC: 7 minutes

Application: Video background modeling

Example: Changes in illumination

Specs

- 1.5 minutes of lobby surveillance (Li, Huang, Gu, and Tian, 2004)
- 1546 frames, 20480 pixels
- Base algorithm: 1.5 hours
- DFC: 8 minutes

Future Directions

New Applications and Datasets

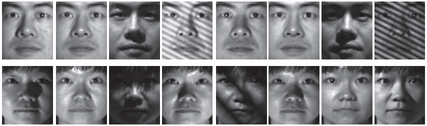
- Practical problems with large-scale or real-time requirements

Example: Large-scale Affinity Estimation

Goal: Estimate semantic similarity between pairs of datapoints

- Motivation: Assign class labels to datapoints based on similarity

Examples from computer vision

- Image tagging: `tree` vs. `firefighter` vs. `Tony Blair`
- Video / multimedia content detection: `wedding` vs. `concert`
- Face clustering: 

Application: Content detection, 9K YouTube videos, 20 classes

- Baseline: Low Rank Representation (Liu, Lin, and Yu, 2010)
  - Strong guarantees but 1.5 days to run
- Divide and conquer (Talwalkar, Mackey, Mu, Chang, and Jordan, 2013)
  - Comparable guarantees
  - Comparable performance in 1 hour (5 subproblems)

Future Directions

New Applications and Datasets

- Practical problems with large-scale or real-time requirements

New Divide-and-Conquer Strategies

- Other ways to reduce computation while preserving accuracy

DFC-NYS: Generalized Nyström Decomposition

1. Choose a random column submatrix $\mathbf{C} \in \mathbb{R}^{m\times l}$ and a random row submatrix $\mathbf{R} \in \mathbb{R}^{d\times n}$ from $\mathbf{M}$. Call their intersection $\mathbf{W}$.

$$\mathbf{M} = \begin{bmatrix} \mathbf{W} & \mathbf{M}_{12} \\ \mathbf{M}_{21} & \mathbf{M}_{22} \end{bmatrix} \quad \mathbf{C} = \begin{bmatrix} \mathbf{W} \\ \mathbf{M}_{21} \end{bmatrix} \quad \mathbf{R} = [\mathbf{W} \quad \mathbf{M}_{12}]$$

2. Recover the low rank components of $\mathbf{C}$ and $\mathbf{R}$ **in parallel** to obtain $\hat{\mathbf{C}}$ and $\hat{\mathbf{R}}$
3. Recover $\mathbf{L}_0$ from $\hat{\mathbf{C}}$, $\hat{\mathbf{R}}$, and their intersection $\hat{\mathbf{W}}$

$$\hat{\mathbf{L}}^{nys} = \hat{\mathbf{C}}\hat{\mathbf{W}}^{+}\hat{\mathbf{R}}$$

   - Generalized Nyström method (Goreinov, Tyrtyshnikov, and Zamarashkin, 1997)
   - Minimal cost: $O(mk^2 + lk^2 + dk^2)$ where $k = \operatorname{rank}(\hat{\mathbf{L}}^{nys})$
4. **Ensemble:** Run p times in parallel and average estimates

Future Directions

New Applications and Datasets

- Practical problems with large-scale or real-time requirements

New Divide-and-Conquer Strategies

- Other ways to reduce computation while preserving accuracy
- More extensive use of ensembling

New Theory

- Analyze statistical implications of divide and conquer algorithms
  - Trade-off between statistical and computational efficiency
  - Impact of ensembling
- Developing suite of matrix concentration inequalities to aid in the analysis of randomized algorithms with matrix data

Concentration Inequalities

Matrix concentration

$$\mathbb{P}\{\|\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}\| \geq t\} \leq \delta$$
$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \geq t\} \leq \delta$$

- Non-asymptotic control of random matrices with complex distributions

Applications

- Matrix completion from sparse random measurements

  (Gross, 2011; Recht, 2011; Negahban and Wainwright, 2010; Mackey, Talwalkar, and Jordan, 2014b)
- Randomized matrix multiplication and factorization

  (Drineas, Mahoney, and Muthukrishnan, 2008; Hsu, Kakade, and Zhang, 2011)
- Convex relaxation of robust or chance-constrained optimization

  (Nemirovski, 2007; So, 2011; Cheung, So, and Wang, 2011)
- Random graph analysis (Christofides and Markström, 2008; Oliveira, 2009)

Concentration Inequalities

Matrix concentration

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{X} - \mathbb{E}\,\boldsymbol{X}) \geq t\} \leq \delta$$

Difficulty: Matrix multiplication is not commutative

$$\Rightarrow \mathrm{e}^{\boldsymbol{X}+\boldsymbol{Y}} \neq \mathrm{e}^{\boldsymbol{X}}\mathrm{e}^{\boldsymbol{Y}} \neq \mathrm{e}^{\boldsymbol{Y}}\mathrm{e}^{\boldsymbol{X}}$$

Past approaches (Ahlswede and Winter, 2002; Oliveira, 2009; Tropp, 2011)

- Rely on deep results from matrix analysis
- Apply to sums of independent matrices and matrix martingales

Our work (Mackey, Jordan, Chen, Farrell, and Tropp, 2014a; Paulin, Mackey, and Tropp, 2015)

- Stein's method of exchangeable pairs (1972), as advanced by Chatterjee (2007) for scalar concentration
  - $\Rightarrow$ Improved exponential tail inequalities (Hoeffding, Bernstein, Bounded differences)
  - $\Rightarrow$ Polynomial moment inequalities (Khintchine, Rosenthal)
  - $\Rightarrow$ Dependent sums and more general matrix functionals

Example: Matrix Bounded Differences Inequality

Corollary (Paulin, Mackey, and Tropp, 2015)

Suppose $Z = (Z_1, \ldots, Z_n)$ has independent coordinates, and

$$\left(\boldsymbol{H}(z_1, \ldots, z_j, \ldots, z_n) - \boldsymbol{H}(z_1, \ldots, z_j', \ldots, z_n)\right)^2 \preccurlyeq \boldsymbol{A}_j^2$$

for all j and values $z_1, \ldots, z_n, z_j'$. Define the boundedness parameter

$$\sigma^2 := \left\| \sum_{j=1}^n \boldsymbol{A}_j^2 \right\|.$$

If each $\boldsymbol{A}_j$ is $d \times d$, then, for all $t \geq 0$,

$$\mathbb{P}\{\lambda_{\max}(\boldsymbol{H}(Z) - \mathbb{E}\,\boldsymbol{H}(Z)) \geq t\} \leq d \cdot \mathrm{e}^{-t^2/(2\sigma^2)}.$$

- Improves prior results in the literature (e.g., Tropp, 2011)
- Useful for analyzing
  - Multiclass classifier performance (Machart and Ralaivola, 2012)
  - Crowdsourcing accuracy (Dalvi, Dasgupta, Kumar, and Rastogi, 2013)
  - Convergence in non-differentiable optimization (Zhou and Hu, 2014)

Future Directions

New Applications and Datasets

- Practical problems with large-scale or real-time requirements

New Divide-and-Conquer Strategies

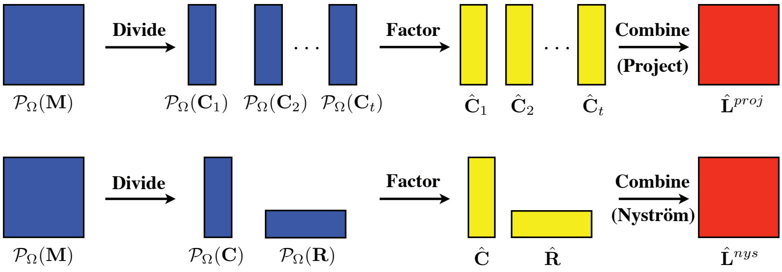
- Other ways to reduce computation while preserving accuracy
- More extensive use of ensembling

New Theory

- Analyze statistical implications of divide and conquer algorithms
  - Trade-off between statistical and computational efficiency
  - Impact of ensembling
- Developing suite of matrix concentration inequalities to aid in the analysis of randomized algorithms with matrix data

The End

Thanks!



References I

Agarwal, A., Negahban, S., and Wainwright, M. J. Noisy matrix decomposition via convex relaxation: Optimal rates in high dimensions. In *International Conference on Machine Learning*, 2011.

Ahlswede, R. and Winter, A. Strong converse for identification via quantum channels. *IEEE Trans. Inform. Theory*, 48(3): 569–579, Mar. 2002.

Cai, J. F., Candès, E. J., and Shen, Z. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4), 2010.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9 (6):717–772, 2009.

Candès, E. J., Li, X., Ma, Y., and Wright, J. Robust principal component analysis? *Journal of the ACM*, 58(3):1–37, 2011.

Candès, E.J. and Plan, Y. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925 –936, 2010.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A., and Willsky, A. S. Sparse and low-rank matrix decompositions. In *Allerton Conference on Communication, Control, and Computing*, 2009.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. preprint, 2010.

Chatterjee, S. Stein's method for concentration inequalities. *Probab. Theory Related Fields*, 138:305–321, 2007.

Cheung, S.-S., So, A. Man-Cho, and Wang, K. Chance-constrained linear matrix inequalities with dependent perturbations: a safe tractable approximation approach. Available at `http://www.optimization-online.org/DB_FILE/2011/01/2898.pdf`, 2011.

Christofides, D. and Markström, K. Expansion properties of random cayley graphs and vertex transitive graphs via matrix martingales. *Random Struct. Algorithms*, 32(1):88–100, 2008.

Dalvi, N., Dasgupta, A., Kumar, R., and Rastogi, V. Aggregating crowdsourced binary ratings. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pp. 285–294, Republic and Canton of Geneva, Switzerland, 2013.

Drineas, P., Mahoney, M. W., and Muthukrishnan, S. Relative-error CUR matrix decompositions. *SIAM Journal on Matrix Analysis and Applications*, 30:844–881, 2008.

Fazel, M., Hindi, H., and Boyd, S. P. A rank minimization heuristic with application to minimum order system approximation. In *In Proceedings of the 2001 American Control Conference*, pp. 4734–4739, 2001.

References II

Frieze, A., Kannan, R., and Vempala, S. Fast Monte-Carlo algorithms for finding low-rank approximations. In *Foundations of Computer Science*, 1998.

Goreinov, S. A., Tyrtyshnikov, E. E., and Zamarashkin, N. L. A theory of pseudoskeleton approximations. *Linear Algebra and its Applications*, 261(1-3):1 – 21, 1997.

Gross, D. Recovering low-rank matrices from few coefficients in any basis. *IEEE Trans. Inform. Theory*, 57(3):1548–1566, Mar. 2011.

Hsu, D., Kakade, S. M., and Zhang, T. Dimension-free tail inequalities for sums of random matrices. Available at `arXiv:1104.1672`, 2011.

Keshavan, R. H., Montanari, A., and Oh, S. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 99: 2057–2078, 2010.

Li, L., Huang, W., Gu, I. Y. H., and Tian, Q. Statistical modeling of complex backgrounds for foreground object detection. *IEEE Transactions on Image Processing*, 13(11):1459–1472, 2004.

Lin, Z., Chen, M., Wu, L., and Ma, Y. The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices. UIUC Technical Report UILU-ENG-09-2215, 2009a.

Lin, Z., Ganesh, A., Wright, J., Wu, L., Chen, M., and Ma, Y. Fast convex optimization algorithms for exact recovery of a corrupted low-rank matrix. UIUC Technical Report UILU-ENG-09-2214, 2009b.

Liu, G., Lin, Z., and Yu, Y. Robust subspace segmentation by low-rank representation. In *International Conference on Machine Learning*, 2010.

Machart, P. and Ralaivola, L. Confusion Matrix Stability Bounds for Multiclass Classification. Available at `http://arXiv.org/abs/1202.6221`, February 2012.

Mackey, L., Jordan, M. I., Chen, R. Y., Farrell, B., and Tropp, J. A. Matrix concentration inequalities via the method of exchangeable pairs. *The Annals of Probability*, 42(3):906–945, 2014a.

Mackey, L., Talwalkar, A., and Jordan, M. I. Distributed matrix completion and robust factorization. *Journal of Machine Learning Research*, 2014b. In press.

References III

Min, K., Zhang, Z., Wright, J., and Ma, Y. Decomposing background topics from keywords by principal component pursuit. In *Conference on Information and Knowledge Management*, 2010.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. arXiv:1009.2118v2[cs.IT], 2010.

Nemirovski, A. Sums of random symmetric matrices and quadratic optimization under orthogonality constraints. *Math. Program.*, 109:283–317, January 2007. ISSN 0025-5610. doi: 10.1007/s10107-006-0033-0. URL http://dl.acm.org/citation.cfm?id=1229716.1229726.

Oliveira, R. I. Concentration of the adjacency matrix and of the Laplacian in random graphs with independent edges. Available at arXiv:0911.0600, Nov. 2009.

Paulin, D., Mackey, L., and Tropp, J. A. Efron-Stein Inequalities for Random Matrices. *The Annals of Probability*, to appear 2015.

Recht, B. Simpler approach to matrix completion. *J. Mach. Learn. Res.*, 12:3413–3430, 2011.

So, A. Man-Cho. Moment inequalities for sums of random matrices and their applications in optimization. *Math. Program.*, 130 (1):125–151, 2011.

Stein, C. A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. 6th Berkeley Symp. Math. Statist. Probab.*, Berkeley, 1972. Univ. California Press.

Talwalkar, Ameet, Mackey, Lester, Mu, Yadong, Chang, Shih-Fu, and Jordan, Michael I. Distributed low-rank subspace segmentation. December 2013.

Toh, K. and Yun, S. An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems. *Pacific Journal of Optimization*, 6(3):615–640, 2010.

Tropp, J. A. User-friendly tail bounds for sums of random matrices. *Found. Comput. Math.*, August 2011.

Zhou, Enlu and Hu, Jiaqiao. Gradient-based adaptive stochastic search for non-differentiable optimization. *Automatic Control, IEEE Transactions on*, 59(7):1818–1832, 2014.

Zhou, Z., Li, X., Wright, J., Candès, E. J., and Ma, Y. Stable principal component pursuit. In *IEEE International Symposium on Information Theory Proceedings (ISIT)*, pp. 1518 –1522, 2010.