An Accelerated Gradient Method for Trace Norm Minimization

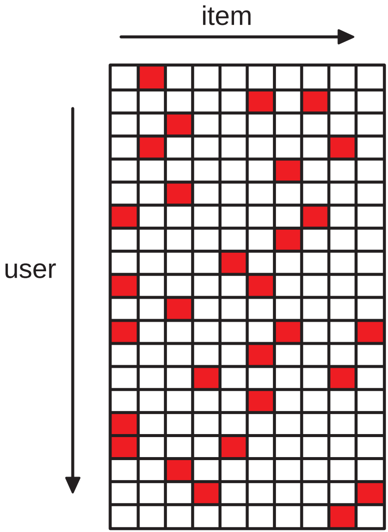
Shuiwang Ji

Arizona State University

ICML, Montreal, June 16th, 2009

Joint work with Jieping Ye

A Motivating Example



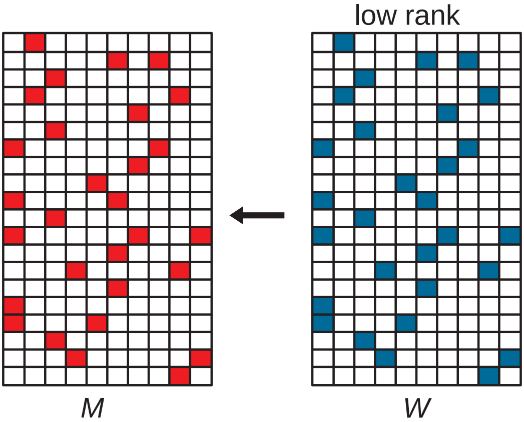
- Give partial rankings of items by some users
- Predict the missing rankings
- A large user-item matrix is given
- Predict the missing entries in the user-item matrix

A matrix completion problem

A Motivating Example–Contd.

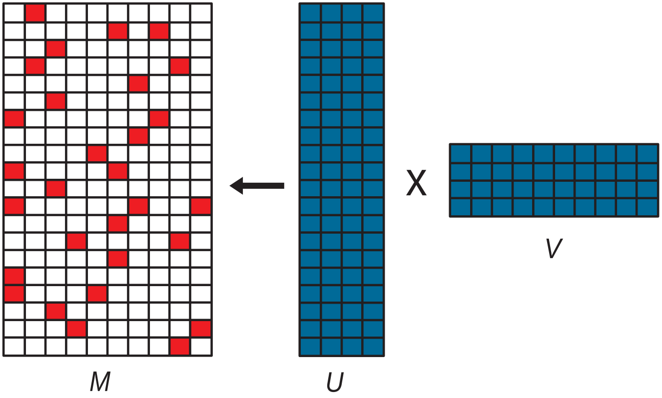
- Only a few factors contribute to a user's taste
- Approximate the rating matrix with a low-rank matrix

$$\min_{W} \sum_{i,j \in \text{observed}} \ell(M_{ij}, W_{ij}) + \lambda * \text{rank}(W)$$



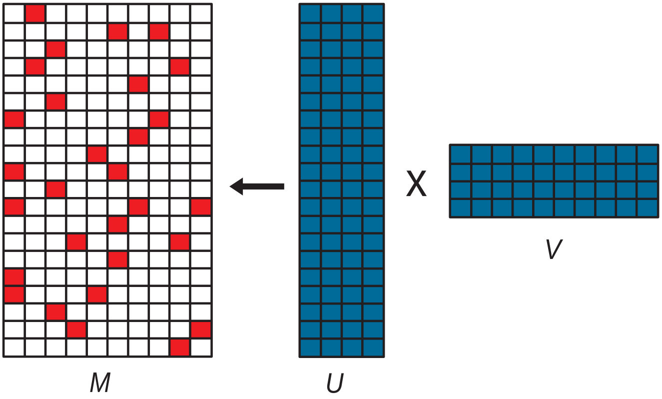
A Motivating Example–Contd.

- Rank minimization is NP-hard
- Assume $W = UV$
- Optimize over U and V iteratively
- Solution is locally optimal



A Motivating Example–Contd.

- Rank minimization is NP-hard
- Assume $W = UV$
- Optimize over U and V iteratively
- Solution is locally optimal



Convex Relaxation of Rank Function

- Trace norm is the convex envelope of the rank function over the unit ball of spectral norm ⇒ a convex relaxation
- Trace norm of a matrix is the sum of its singular values:

$$W = U \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma_k \end{pmatrix} V^T$$

$$||W||_* = \sum_{i=1}^{k} \sigma_i = ||(\sigma_1, \cdots, \sigma_k)||_1$$

- trace norm ≈ rank ⇔ $L_1 \approx L_0$

Problem Formulation

$$\min_{W} F(W) = \underbrace{f(W)}_{\text{loss}} + \underbrace{\lambda ||W||_*}_{\text{regularization}}$$

- $W \in \mathbb{R}^{m \times n}$: the matrix variable
- The gradient of $f(\cdot)$ is Lipschitz continuous:

$$|| \bigtriangledown f(X) - \bigtriangledown f(Y)||_F \leq L||X - Y||_F, \forall X, Y \in \mathbb{R}^{m \times n}$$

- $||W||_*$ is NOT a smooth (differentiable) function

Trace Norm Regularized Problems

- **Matrix completion** (Srebro *et al.* 2005, Candés & Recht, 2008): $f(W) = \sum_{(i,j)\in\Omega} \ell(M_{ij}, W_{ij})$
  - $M \in \mathbb{R}^{m\times n}$: the partially observed matrix with the entries in Ω being observed

$$\min_W \sum_{i,j\in\Omega} \ell(M_{ij}, W_{ij}) + \lambda ||W||_*$$

- **Multi-task learning** (Abernethy *et al.* 2006, Argyriou *et al.* 2008): $f(W) = \sum_{i=1}^{n} \sum_{j=1}^{s_i} \ell(y_i^j, w_i^T x_i^j)$
  - n: the number of tasks
  - $(x_i^j, y_i^j) \in \mathbb{R}^m \times \mathbb{R}$: the jth sample in the ith task
  - s_i: the number of samples in the ith task
  - $W = [w_1, \cdots, w_n] \in \mathbb{R}^{m\times n}$
- **Matrix classification** (Tomioka *et al.* 2008, Bach 2008): $f(W) = \sum_{i=1}^{s} \ell(y_i, \mathrm{Tr}(W^T X_i))$
  - $(X_i, y_i) \in \mathbb{R}^{m\times n} \times \mathbb{R}$: the ith sample

The Subgradient Method

- Trace norm is non-smooth
- Apply the subgradient method as

$$W_k = W_{k-1} - \frac{1}{t_k} \underbrace{F'(W_{k-1})}_{\text{subgradient at } W_{k-1}}$$

- The subgradient method converges as $O(\frac{1}{\sqrt{k}})$:

$$F(W_k) - F(W^*) \leq c\frac{1}{\sqrt{k}}$$

Remark

- *This convergence rate is optimal for non-smooth problems under the first-order black-box model (Nesterov 2003)*
- *Convergence rate cannot be improved if no special structure of the trace norm is exploited*

The Subgradient Method

- Trace norm is non-smooth
- Apply the subgradient method as

$$W_k = W_{k-1} - \frac{1}{t_k} \underbrace{F'(W_{k-1})}_{\text{subgradient at } W_{k-1}}$$

- The subgradient method converges as $O(\frac{1}{\sqrt{k}})$:

$$F(W_k) - F(W^*) \leq c\frac{1}{\sqrt{k}}$$

Remark

- *This convergence rate is optimal for non-smooth problems under the first-order black-box model (Nesterov 2003)*
- *Convergence rate cannot be improved if no special structure of the trace norm is exploited*

Our Main Contributions

By exploiting the special structures of trace norm, we propose two algorithms:

- Extended Gradient Algorithm: converges as $O(\frac{1}{k})$
- Accelerated Gradient Algorithm: converges as $O(\frac{1}{k^2})$

Remark

$O(\frac{1}{k^2})$ is the optimal convergence rate for smooth problems (Nesterov 2003) $\Rightarrow$ the non-smoothness effect of trace norm is removed

Our Main Contributions

By exploiting the special structures of trace norm, we propose two algorithms:

- Extended Gradient Algorithm: converges as $O(\frac{1}{k})$
- Accelerated Gradient Algorithm: converges as $O(\frac{1}{k^2})$

Remark

$O(\frac{1}{k^2})$ is the optimal convergence rate for smooth problems (Nesterov 2003) $\Rightarrow$ the non-smoothness effect of trace norm is removed

Two Equivalent Views of Gradient Descent

- Consider the minimization of the smooth function

$$\min_{W} f(W)$$

using gradient descent:

$$W_k = W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1})$$

- It can be reformulated equivalently as

$$W_k = \arg\min_{W} \left\{ \underbrace{f(W_{k-1}) + \langle W - W_{k-1}, \bigtriangledown f(W_{k-1}) \rangle}_{\text{linear approximation at } W_{k-1}} + \underbrace{\frac{t_k}{2} ||W - W_{k-1}||_F^2}_{\text{regularization}} \right\}$$

- What about $||W||_*$?

Two Equivalent Views of Gradient Descent

- Consider the minimization of the smooth function

$$\min_{W} f(W)$$

using gradient descent:

$$W_k = W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1})$$

- It can be reformulated equivalently as

$$W_k = \arg\min_{W} \left\{ \underbrace{f(W_{k-1}) + \langle W - W_{k-1}, \bigtriangledown f(W_{k-1}) \rangle}_{\text{linear approximation at } W_{k-1}} + \underbrace{\frac{t_k}{2} ||W - W_{k-1}||_F^2}_{\text{regularization}} \right\}$$

- What about $||W||_*$?

Incorporating the Non-smooth Term

- Add $\lambda||W||_*$ directly without approximation
- Solve the trace norm regularized problem by the iterative step:

$$W_k = \arg\min_W \left\{ \underbrace{\text{linear approximation} + \text{regularization}}_{\text{corresponds to } f(W)} + \lambda||W||_* \right\}$$

- It can be expressed equivalently as

$$W_k = \arg\min_W \left\{ \frac{t_k}{2} ||W - A||_F^2 + \lambda ||W||_* \right\}$$

where $A = W_{k-1} - \frac{1}{t_k} \bigtriangledown f(W_{k-1})$

- The above problem can be solved by first computing the SVD of A and then applying soft thresholding on the singular values

A Key Theorem

Theorem

Let $C = U\Sigma V^T$ be the SVD of C. Then

$$\mathcal{T}_\lambda(C) \equiv \arg\min_W \left\{ \frac{1}{2}||W - C||_F^2 + \lambda ||W||_* \right\}$$

is given by

$$\mathcal{T}_\lambda(C) = U\Sigma_\lambda V^T,$$

where Σ_λ is diagonal with

$$(\Sigma_\lambda)_{ii} = \underbrace{\max\{0, \Sigma_{ii} - \lambda\}}_{\textit{soft thresholding}}.$$

The Extended Gradient Algorithm

- Initialize $W_0 \in \mathbb{R}^{m \times n}$
- Iterate:
  1. Choose an appropriate step size s_k
  2. Gradient descent: $\tilde{W}_k = W_{k-1} - s_k \bigtriangledown f(W_{k-1})$
  3. Soft thresholding: $W_k = \mathcal{T}_\lambda(\tilde{W}_k)$

The Step Size

- Start from an initial value, decrease by a multiplicative factor $\gamma < 1$, until a condition is satisfied
- If $s_k < \frac{1}{L} \Rightarrow$ the condition is satisfied
- At step t, we use s_{t-1} as initial value

Convergence Analysis

Theorem

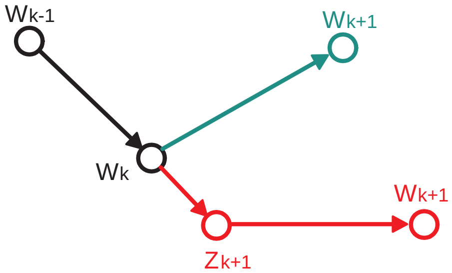
Let $\{W_k\}$ be the sequence generated by the Extended Gradient Algorithm. Then for any $k \geq 1$ we have

$$F(W_k) - F(W^*) \leq \frac{\gamma L ||W_0 - W^*||_F^2}{2k} = O(\frac{1}{k}),$$

where $W^* = \arg\min_W F(W)$.

Nesterov's Acceleration Technique

- The convergence rate of gradient descent for smooth problems is not optimal
- The optimal convergence rate can be achieved by the Nesterov's extrapolation technique (Nesterov 1983, Nesterov 2003)
  - Define two sequences W_k and Z_k
  - Z_{k+1} is affine combination of W_k and W_{k-1}
  - Perform gradient descent at Z_{k+1} instead of W_k



The Accelerated Gradient Algorithm

- Initialize $W_0, Z_1 \in \mathbb{R}^{m \times n}, \alpha_1 = 1$
- Iterate:
  1. Choose an appropriate step size s_k
  2. Gradient descent: $\tilde{W}_k = Z_k - s_k \bigtriangledown f(Z_k)$
  3. Soft thresholding: $W_k = \mathcal{T}_\lambda(\tilde{W}_k)$
  4. $\alpha_{k+1} = \frac{1+\sqrt{1+4\alpha_k^2}}{2}$ compute coefficient
  5. $Z_{k+1} = W_k + \left(\frac{\alpha_k - 1}{\alpha_{k+1}}\right)(W_k - W_{k-1})$ extrapolation

Convergence Analysis

Theorem

Let $\{W_k\}$ and $\{Z_k\}$ be the sequences generated by the Accelerated Gradient Algorithm. Then for any $k \geq 1$ we have

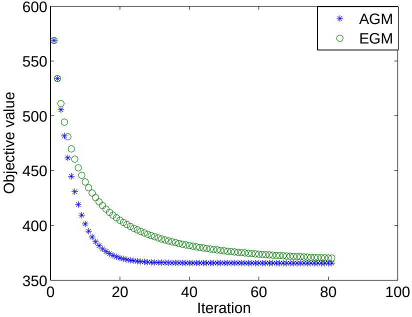
$$F(W_k) - F(W^*) \leq \frac{2\gamma L||W_0 - W^*||_F^2}{(k+1)^2} = O(\frac{1}{k^2}).$$

Evaluation of Efficiency

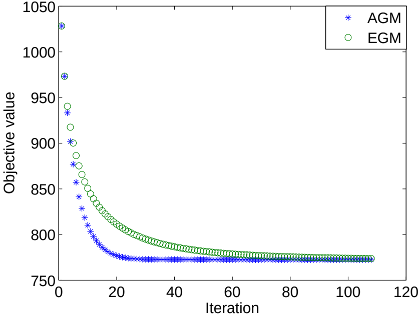
- Use multi-task formulation for evaluation
  - Extended Gradient Method (EGM)
  - Accelerated Gradient Method (AGM)
  - Multi-task Feature Learning (MFL) (Argyriou *et al.* 2008)

Data set	yeast		letters		digits		dmoz	
Percentage	5%	10%	5%	10%	5%	10%	5%	10%
EGM	2.24	3.37	4.74	5.67	62.51	29.59	133.21	146.58
AGM	0.34	0.49	0.62	0.91	2.41	2.39	1.59	1.42
MFL	2.33	17.27	2.49	9.66	15.50	42.64	74.24	31.49

Evaluation of Convergence



yeast (5%)



yeast (10%)

Conclusion and Discussion

- Propose two algorithms for solving trace norm regularized problems
  - Extended Gradient Method
  - Accelerated Gradient Method

$$O(\frac{1}{\sqrt{k}}) \Rightarrow O(\frac{1}{k}) \Rightarrow O(\frac{1}{k^2})$$

- Future work:
  - Approximate SVD to reduce computational cost
  - Adapt the algorithms to constrained problems:

$$\begin{aligned} \min \quad & ||W||_* \\ \text{s.t.} \quad & \text{affine constraints} \end{aligned}$$

Thank you!