A Proximal Point Method for Matrix Least Squares Problem with Nuclear Norm Regularization

Defeng Sun

Department of Mathematics

National University of Singapore

May 28, 2009

Joint work with Kaifeng Jiang and Kim Chuan Toh

Let $\mathcal{S}^n$ be the set of all real symmetric matrices and $\mathcal{S}^n_+$ be the cone of all positive semidefinite matrices in $\mathcal{S}^n$.

We consider the least squares SDP:

$$\min\left\{\frac{1}{2}\|\mathcal{A}(X)-b\|^2+\rho\langle I,\,X\rangle\;:\;\mathcal{B}(X)=d,\;X\in\mathcal{S}^n_+\right\},$$

where $\mathcal{A}:\mathcal{S}^n\to\Re^m$ and $\mathcal{B}:\mathcal{S}^n\to\Re^s$ are linear maps and ρ is a given positive scalar.

Difficulty: even $\mathcal{A}=\mathcal{I}$, the problem can be difficult to solve.

An example — the regularized kernel estimation (RKE) problem in statistics:

we are given a set of n objects and dissimilarity measures d_{ij} for certain object pairs $(i, j) \in \mathcal{E}$.

The goal is to estimate a positive semidefinite kernel matrix $X \in \mathcal{S}^n_+$ such that the fitted squared distances between objects induced by X satisfy

$$X_{ii} + X_{jj} - 2X_{ij} = \langle A_{ij}, X \rangle \approx d_{ij}^2 \quad \forall\, (i, j) \in \mathcal{E},$$

where $A_{ij} = (e_i - e_j)(e_i - e_j)^T$.

One version of the RKE problem is to solve the following SDP:

$$\min\Big\{ \sum_{(i,j)\in\mathcal{E}} W_{ij}(\langle A_{ij},\, X\rangle - d_{ij}^2)^2 + \rho\langle I,\, X\rangle \ : \ \langle E,\, X\rangle = 0,\ X \succeq 0\Big\},$$

where $W \in \mathcal{S}^n$ is a given weight matrix with positive entries.

Analogously, we consider the least squares problem with the nuclear norm regularization:

$$\min\left\{\frac{1}{2}\|\mathcal{A}(X)-b\|^2+\rho\|X\|_* \,:\, \mathcal{B}(X)=d,\ X\in\Re^{p\times q}\right\},$$

where

$$\|X\|_*=\sum_{i=1}^{k}\sigma_i(X)$$

and $\sigma_i(X)$ are the singular values of X.

The matrix completion example:

$$\min \left\{ \mathrm{rank}(X) \;:\; X_{ij} \approx M_{ij} \;\; \forall\, (i,j) \in \Omega \right\},$$

where

$$\Omega \in \{1,\ldots,p\} \times \{1,\ldots,q\} :$$

$$\begin{bmatrix} * & & & & * & \\ & * & & & & * \\ * & & & & * & \\ & & * & * & & \end{bmatrix}$$

get a relaxed convex problem:

$$\min\left\{\|X\|_* \;:\; X_{ij} \approx M_{ij} \;\; \forall\,(i,j) \in \Omega\right\}.$$

Further

$$\min\left\{\frac{1}{2}\sum_{(i,j)\in\Omega}(X_{ij} - M_{ij})^2 + \rho\|X\|_*\right\}.$$

The Netflix Prize problem: the convex relaxation is pretty good.

http://www.netflixprize.com/index

For a random example:

- $p = q = 10^5$, $\text{rank}(X) = 10$, noise level $=0.1$.
- $|\Omega| \approx 1.2 \times 10^7$.
- Proximal point method framework + gradient projection method.
- Need 416 seconds to achieve a relative accuracy 0.0453.

Consider the Moreau-Yosida regularization:

$$\begin{array}{rll} F_\sigma(X) = & \min & \frac{1}{2}\|u\|^2 + \rho\|Y\|_* + \frac{1}{2\sigma}\|Y - X\|^2 \\ & \text{s.t.} & \mathcal{A}(Y) + u \ = \ b \\ & & \mathcal{B}(Y) \qquad\quad = \ d \\ & & Y \in \Re^{p\times q}, \quad u \in \Re^m. \end{array} \tag{1}$$

The Lagrangian dual problem of (1) is

$$\max_{y\in\Re^m, z\in\Re^s} \Big\{ \theta^\rho_\sigma(y,z;X) := \inf_{u\in\Re^m, Y\in\Re^{p\times q}} L^\rho_\sigma(Y,u;y,z,X)$$

$$= -\frac{1}{2}\|y\|^2 + \langle b, y\rangle + \langle d, z\rangle$$

$$+\frac{1}{2\sigma}\|X\|^2 - \frac{1}{2\sigma}\|D_{\rho\sigma}(W(y,z;X))\|^2 \Big\}, \quad (2)$$

where $W(y,z;X) = X + \sigma(\mathcal{A}^* y + \mathcal{B}^* z)$.

For any $Y \in \Re^{p\times q}$, $D_\rho(Y)$ is the unique optimal solution to the following strongly convex function

$$\min_X \|X\|_* + \frac{1}{2\rho}\|X - Y\|_F^2$$

It is well known that $D_\rho(\cdot)$ is globally Lispchitz continuous with modulus 1.

Let $Y \in \Re^{p\times q}$ admit the following singular value decomposition:

$$Y = U[\Sigma \ 0]V^T,$$

where $U \in \Re^{p\times p}$ and $V \in \Re^{q\times q}$ are orthogonal matrices, $\Sigma = \mathrm{diag}(\sigma_1, \cdots, \sigma_p)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are singular values of Y. For each $\rho > 0$, the operator D_ρ is given by:

$$D_\rho(Y) = U[\Sigma_\rho \ 0]V^T,$$

where $\Sigma_\rho = \mathrm{diag}((\sigma_1 - \rho)_+, \ldots, (\sigma_p - \rho)_+)$.

Good news is: $\|D_\rho(Y)\|^2$ is continuously differentiable and

$$\nabla\Big(\frac{1}{2}\|D_\rho(Y)\|^2\Big) = D_\rho(Y).$$

So,

$$\begin{aligned}\theta_\sigma^\rho(y,z;X) &= -\frac{1}{2}\|y\|^2 + \langle b, y\rangle + \langle d, z\rangle \\ &\quad + \frac{1}{2\sigma}\|X\|^2 - \frac{1}{2\sigma}\|D_{\rho\sigma}(W(y,z;X))\|^2,\end{aligned}$$

is continuously differentiable, where

$$W(y,z;X) = X + \sigma(\mathcal{A}^* y + \mathcal{B}^* z).$$

The Moreau-Yosida regularization:

$$\begin{array}{rll} F_\sigma(X) = & \min & \frac{1}{2}\|u\|^2 + \rho\|Y\|_* + \frac{1}{2\sigma}\|Y - X\|^2 \\ & \text{s.t.} & \mathcal{A}(Y) + u \ = \ b \\ & & \mathcal{B}(Y) \qquad\quad = \ d \\ & & Y \in \Re^{p\times q}, \quad u \in \Re^m \end{array}$$

$$\Downarrow$$

a smooth convex optimization problem:

$$\max_{y\in\Re^m, z\in\Re^s} \left\{ \theta_\sigma^\rho(y, z; X) \right\}.$$

Even better: $D_\rho(\cdot)$ is **strongly semismooth** everywhere.

A Lipschitz function $F : \mathcal{X} \to \mathcal{Y}$ is said to be strongly semismooth at $x \in \mathcal{X}$ if

1) it is directionally differentiable at x; and 2)

$$F(x + \Delta x) - F(x) - F'(x + \Delta x)\Delta x = O(\|\Delta x\|^2)$$

for all $x + \Delta x$ such that F is Fréchet differentiable at $x + \Delta x$.

One key issue:

$$\theta_\sigma^\rho(\cdot,\cdot;X) \notin \mathcal{C}^2.$$

This property allows $\theta_\sigma^\rho(\cdot,\cdot;X)$ to possess
<u>negative definite (generalized) Hessian</u>,

which is vital for an inexact second order method to be efficient.

We apply the **proximal point method** to solve

$$\min_{X\in\Re^{p\times q}} \left\{\Phi_\sigma^\rho(X) := \max\{\theta_\sigma^\rho(y,z;X) \,:\, y\in\Re^m,\ z\in\Re^s\}\right\}.$$

$$\Uparrow$$

$\theta_\sigma^\rho(y,z;X)$ `via the dual of the MY regularization`

$$\min\left\{\frac{1}{2}\|\mathcal{A}(X)-b\|^2+\rho\|X\|_* \,:\, \mathcal{B}(X)=d,\ X\in\Re^{p\times q}\right\}.$$

PPA. Input $X^0 \in \Re^{p\times q}$, $\sigma_0 > 0$, iterate:

1. Compute an approximate maximizer

$$(y^k, z^k) \;\approx\; \mathrm{argmax}\{\theta^{\rho}_{\sigma_k}(y, z; X^k) \,:\, y \in \Re^m,\; z \in \Re^s\},$$

2. $X^{k+1} \;=\; D_{\rho\sigma_k}(W(y^k, z^k; X^k)), \quad Z^{k+1} = \dfrac{1}{\sigma_k}(D_{\rho\sigma_k}(W(y^k, z^k; X^k)) - W(y^k, z^k; X^k)),$

3. If $\|R_d^k := \mathcal{A}^* y^k + \mathcal{B}^* z^k + Z^{k+1}\|_F \leq \varepsilon$; stop; else, update σ_k.

For the inner subproblem, the optimality condition is given by

$$\begin{aligned} \nabla_y \theta^\rho_{\sigma_k}(y, z; X^k) &= b - y - \mathcal{A} D_{\rho\sigma}(W(y, z; X^k)) = 0 \\ \nabla_z \theta^\rho_{\sigma_k}(y, z; X^k) &= d - \mathcal{B} D_{\rho\sigma}(W(y, z; X^k)) = 0 \end{aligned} \tag{3}$$

We solve (3) by a **semismooth Newton-CG method.**

The inner problems can be solved by a **(fast) semismooth Newton-CG method**. The outer iteration

$$X^{k+1} \;=\; D_{\rho\sigma_k}(W(y^k, z^k; X^k))$$

only satisfies

$$X^{k+1} = X^k - \sigma_k \nabla \Phi^{\rho}_{\sigma_k}(X^k),$$

a gradient descent step. The exciting news is that it can also be seen as

<u>**an approximate semismooth Newton method**</u>, at least for the least squares SDP case.

Selected examples:

1. For each pair (n, r), we generate a positive semidefinite matrix $M \in \mathcal{S}^n$ of rank r by setting $M = M_1 M_1^T$ where $M_1 \in \Re^{n\times r}$ is a random matrix with i.i.d Gaussian entries. Then we sample a subset Ω of m entries uniformly at random from the upper triangular part of M. The observed data is set to be $\widetilde{M}_\Omega = M_\Omega + \alpha N_\Omega \|M_\Omega\|_F / \|N_\Omega\|_F$, where the random matrix $N_\Omega \in \mathcal{S}^n$ is generated that has sparsity pattern Ω and i.i.d Gaussian entries and α is the noise level.

The minimization problem we solve is given by

$$\min\left\{\frac{1}{2}\|X_\Omega - \widetilde{M}_\Omega\|_F^2 + \rho\langle I,\, X\rangle \;:\; X \succeq 0 \;\right\}. \tag{4}$$

Numerical results: $n = 2000$, $r = 100$,

– for $\alpha = 0$, we need 15:00 and 8 (27) iterations; and

– for $\alpha = 0.05$, we need 39:15 and 18(63) iterations

– The relative accuracy is below 10^{-6}.

– The averaged CGs each step ≤ 10.

– $|\Omega| \approx 975,000$.

2. The nonsymmtric problem: similarly generated as in Example 1.

Numerical results: $p = q = 1000$, $r = 50$,

– for $\alpha = 0$, we need **4:07** and **12** (24) iterations; and

– for $\alpha = 0.05$, we need **16:01** and **26** (73) iterations.

– The averaged CGs each step ≤ 5.

– The relative accuracy is below 10^{-6}.

– $|\Omega| = 487,500$.