A Proximal Point Algorithm for Nuclear Norm Regularized Matrix Least Squares Problems

Kim-Chuan Toh

Department of Mathematics, National University of Singapore

Workshop on Optimization, Fields Institute

Joint work with Kaifeng Jiang and Defeng Sun

Outline

1. Introduction
2. Nuclear norm regularized LS problems with linear equality/inequality constraints
3. An inexact partial proximal point algorithm (PPA)
4. Strong semismoothness of the soft-thresholding operator
5. An inexact smoothing Newton method for solving PPA subproblems
6. Quadratic convergence of inexact smoothing Newton method under constraint nondegeneracy condition
7. Numerical performance
8. Conclusions & future work

Unconstrained nuclear norm regularized LS problem

The affine rank minimization problem has been intensively studied:

$$\min \left\{ \operatorname{rank}(X) \;:\; \mathcal{A}(X) \;=\; b,\; X \in \Re^{p\times q} \right\} \quad \text{(NP-hard)}$$

where $\mathcal{A} : \mathbb{R}^{p\times q} \to \mathbb{R}^m$ is a linear map and $b \in \mathbb{R}^m$. We assume $p \leq q$ w.l.o.g.
[Fazel 2002] considered the nuclear norm convex relaxation:

$$\min \left\{ \|X\|_* = \sum_{i=1}^{p} \sigma_i(X) \;:\; \mathcal{A}(X) \;=\; b,\; X \in \mathbb{R}^{p\times q} \right\}. \tag{1}$$

where $\sigma_i(X)$'s are singular values of X.
For problems with noisy data b, one would typically consider the matrix LS problem with nuclear norm regularization:

$$\min \left\{ \frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\|X\|_* \;:\; X \in \mathbb{R}^{p\times q} \right\}. \tag{2}$$

It is well known that (1) can be reformulated as an SDP:

$$\min \left\{ \frac{1}{2}(\operatorname{Tr}(W_1) + \operatorname{Tr}(W_2)) \;:\; \mathcal{A}(X) = b,\; \begin{pmatrix} W_1 & X \\ X^T & W_2 \end{pmatrix} \succeq 0 \right\}.$$

But state-of-the-art interior-point solvers like SeDuMi or SDPT3 are not suitable for problems with large m or $p + q$. When $p \ll q$, it is especially advantageous to design algorithms which deal with X directly.

Some recent approaches

Problem (1) or (2) arises frequently in matrix completion, dimension reduction in multivariate linear regression, multi-class classification/learning.

1. [Cai,Candès,Shen 2008] designed the SVT algorithm for solving the following Tikhonov regularized version of (1):

$$\min\left\{\|X\|_* + \frac{1}{2\tau}\|X\|^2 \ : \ \mathcal{A}(X) \ = \ b, \ X \in \mathbb{R}^{p\times q}\right\}.$$

2. [Ma,Goldfarb,Chen 2008] developed a fixed point continuation (FPC) method for (2) and a Bregman algorithm for (1).
3. [Toh,Yun 2009] developed an APG algorithm for (2).
4. [Liu,Sun,Toh 2009] developed inexact proximal point algorithms (PPA) for (1) with linear and second order cone constraints.
5. [Pong,Tseng,Ji,Ye 2010] developed APG and PG-type methods for solving various reformulations of the following problem arising from multi-task learning:

$$\min_X\{\|AX - B\|^2 + \rho\|X\|_*\}$$

6. Many papers in recent ICML conferences dealing with some special variants of nuclear norm regularized problems.

Example 1

In many applications, we may want a low rank approximation X to a target matrix M while preserving certain structures, say nonnegative entries (e.g. concentrations, intensity values), or bounds on the entries.

We consider the nearest matrix approximation problem in [Golub,Hoffman,Stewart 87] where the classic Eckart-Young-Mirsky theorem was extended to obtain the nearest lower-rank approximation while certain columns are fixed:

$$\min_{X \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2}\|X - M\|^2 \;\middle|\; Xe_1 = Me_1,\ \mathrm{rank}(X) \le r \right\}.$$

We may consider the same problem but with the added constraints $X \geq 0$:

$$\min_{X \in \mathbb{R}^{p \times q}} \left\{ \frac{1}{2}\|X - M\|^2 + \rho\|X\|_* \;\middle|\; Xe_1 = Me_1,\ X \geq 0 \right\}.$$

For approximation by a stochastic matrix, impose "$Xe = e$".

Example 2

Given the largest positive eigenvalue λ and the left and right principal eigenvectors of M, find a low rank approximation of M while preserving the left and right principal eigenvectors of M [Ho and Van Dooren 2008]. The problem can be stated as follows:

$$\min_{X \in \mathbb{R}^{n\times n}} \left\{ \frac{1}{2}\|X_{\mathcal{E}} - M_{\mathcal{E}}\|^2 + \rho\|X\|_* \,:\, Xv = \lambda v, X^T w = \lambda w, X \geq 0 \right\}.$$

[Bonacich 1972] used the principal eigenvector to measure the network centrality. The Google's PageRank is a variant of the eigenvector centrality for ranking web pages.

Nuclear norm regularized matrix LS problems

We consider the following nuclear norm regularized matrix LS problem with linear equality and inequality constraints:

$$\text{(NNLS)} \qquad \min_{X\in\mathbb{R}^{p\times q}} \Big\{ f_\rho(X) := \frac{1}{2}\|\mathcal{A}(X)-b\|^2 + \langle C,\, X\rangle + \rho\|X\|_* \;\mid\; \mathcal{B}(X) \in d+\mathcal{Q} \Big\},$$

where $\mathcal{B}:\mathbb{R}^{p\times q}\to\mathbb{R}^s$ is a linear map, $d\in\mathbb{R}^s$, $C\in\mathbb{R}^{p\times q}$, and $\mathcal{Q}=\{0\}^{s_1}\times\mathbb{R}^{s_2}_+$ is a convex polyhedral cone.

Let $u = b-\mathcal{A}(X)$. We will study the equivalent problem:

$$\boxed{\min_{u,X} \Big\{ f_\rho(u,X) := \frac{1}{2}\|u\|^2 + \langle C,\, X\rangle + \rho\|X\|_* \;\Big|\; \begin{array}{c} \mathcal{A}(X)+u=b \\ \mathcal{B}(X) \in d+\mathcal{Q} \end{array} \Big\} \qquad (3)}$$

The dual problem of (3) is given by:

$$\max_{\zeta\in\Re^m,\ \xi\in\mathcal{Q}^*} \Big\{ -\frac{1}{2}\|\zeta\|^2 + \langle b,\zeta\rangle + \langle d,\xi\rangle \;\mid\; \mathcal{A}^*(\zeta)+\mathcal{B}^*(\xi)+Z = C,\ \|Z\|_2 \le \rho \Big\}.$$

Why is NNLS useful for rank constrained LS problem?

Consider the following rank constrained LS problem:

$$\min \left\{ \frac{1}{2}\|\mathcal{A}(X) - b\|^2 \;\mid\; \mathcal{B}(X) \in d + \mathcal{Q},\; \mathrm{rank}(X) \le r \right\}$$

By noting that the rank constraint is equivalent to $\sum_{i=r+1}^{p} \sigma_i = 0 = \|X\|_* - \sum_{i=1}^{r} \sigma_i$, we can consider a penalty approach for the above problem, and start with the penalized objective function:

$$\frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\|X\|_* - \rho \sum_{i=1}^{r} \sigma_i(X) \quad \text{(difference of 2 convex functions)}$$

Given X^k, we can majorize the above function by noting that

$$-\sum_{i=1}^{r} \sigma_i(X) \le -\sum_{i=1}^{r} \sigma_i(X^k) - \langle W^k, X - X^k \rangle \quad \forall\, X$$

where W^k is a subgradient of $\sum_{i=1}^{r} \sigma_i(X)$ at X^k. The majorized penalty problem associated with X^k is the following NNLS:

$$\min \left\{ \frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\|X\|_* - \rho\langle W^k, X \rangle \;\mid\; \mathcal{B}(X) \in d + \mathcal{Q} \right\}$$

A partial proximal point algorithm

Given a starting point (u^0, X^0), the inexact partial PPA generates a sequence (u^k, X^k) by approximately solving the following problem [Rockafellar 1976], [Ha 1990], [Ibaraki,Fukushima 1996]:

$$(u^{k+1}, X^{k+1}) \approx \arg\min \left\{ f_\rho(u,X) + \frac{1}{2\sigma_k}\|X - X^k\|^2 \;\middle|\; \begin{array}{l} \mathcal{A}(X) + u = b \\ \mathcal{B}(X) \in d + \mathcal{Q} \end{array} \right\} \quad (4)$$

where $\{\sigma_k > 0\}$ is a given nondecreasing sequence.
Let $\mathcal{F}$ be the feasible region. Define

$$\widehat{f}_\rho(u,X) = \begin{cases} f_\rho(u,X) & (u,X) \in \mathcal{F} \\ +\infty & \text{otherwise.} \end{cases}$$

Then (4) can be compactly written as:

$$\begin{array}{c} (u^{k+1}, X^{k+1}) \approx \arg\min \left\{ \widehat{f}_\rho(u,X) + \frac{1}{2\sigma_k}\|X - X^k\|^2 \right\} \\ \| \\ P_{\sigma_k}(u^k, X^k) := (\Pi + \sigma_k \partial \widehat{f}_\rho)^{-1}\Pi(u^k, X^k) \end{array}$$

where $\Pi(u,X) = (0,X)$ is the projection of $\mathbb{R}^m \times \mathbb{R}^{p\times q}$ onto $\{0_m\} \times \mathbb{R}^{p\times q}$.
In the classical PPA of Rockefellar, we have the identity $\mathcal{I}$ instead of Π. Much of the convergence theory for the classical PPA can be extended to the above setting.

Moreau-Yosida regularization

In each PPA iteration, we need to solve the following subproblem:

$$F_\sigma(X) = \min_{u,Y} \left\{ \frac{1}{2}\|u\|^2 + \langle C,\ Y\rangle + \rho\|Y\|_* + \frac{1}{2\sigma}\|Y - X\|^2 \ \Bigg|\ \begin{array}{l} \mathcal{A}(Y) + u = b \\ \mathcal{B}(Y) \in d + \mathcal{Q} \end{array} \right\} \quad (5)$$

The Lagrangian dual problem of (5) is given by:

$$\sup\{\Theta_\sigma(\zeta, \xi; X) \ |\ \zeta \in \mathbb{R}^m,\ \xi \in \mathcal{Q}^*\} \quad (6)$$

where

$$\Theta_\sigma(\zeta, \xi; X) := -\frac{1}{2}\|\zeta\|^2 + \langle b,\ \zeta\rangle + \langle d,\ \xi\rangle + \frac{1}{2\sigma}\|X\|^2 - \frac{1}{2\sigma}\|D_{\rho\sigma}(W(\zeta, \xi; X))\|^2,$$

$W(\zeta, \xi; X) = X - \sigma(C - \mathcal{A}^*\zeta - \mathcal{B}^*\xi)$.
By the saddle point theorem [Rockafellar 1970], we know that $D_{\rho\sigma}(W(\zeta, \xi; X))$ is the unique solution to (5) for any

$$(\zeta(X), \xi(X)) \in \text{argsup}\{\Theta_\sigma(\zeta, \xi; X) \mid \zeta \in \mathbb{R}^m,\ \xi \in \mathcal{Q}^*\}$$

Soft Thresholding Operator $D_\rho(\cdot)$

Let $(t)_+ = \max\{t, 0\}$. Define the soft thresholding function $g_\rho : \mathbb{R} \to \mathbb{R}$ by

$$g_\rho(t) := (t - \rho)_+ - (-t - \rho)_+$$

Let the SVD of $Y \in \Re^{p \times q}$ be:

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

where $U \in \mathbb{R}^{p \times p}$ and $V \in \mathbb{R}^{q \times q}$ are orthogonal, $\Sigma = \mathrm{Diag}(\sigma_1, \cdots, \sigma_p)$, and $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are singular values arranged in decreasing order.

For any given $Y \in \mathbb{R}^{p \times q}$ and threshold $\rho > 0$,

$$D_\rho(Y) = \mathrm{argmin}_X \left\{ \|X\|_* + \frac{1}{2\rho}\|X - Y\|^2 \right\}.$$

Based on [Lemaréchal,Sagastizábal 97], it is known that $D_\rho(\cdot)$ is globally Lipschitz continuous with modulus 1.

The soft thresholding operator D_ρ is analytically given by

$$D_\rho(Y) = U[g_\rho(\Sigma),\ 0]V^T = U[(\Sigma - \rho I)_+,\ 0]V^T \tag{7}$$

Note: $D_\rho(\cdot)$ is not differentiable everywhere, but $\|D_\rho(\cdot)\|^2$ is continuously differentiable with

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

Strong semismoothness of $D_\rho(\cdot)$

A locally Lipschitz function $F : \Re^m \to \Re^l$ is strongly semismooth at x if

1. F is directionally differentiable at x
2. for any $h \in \Re^m$ and $V \in \partial F(x+h)$ with $h \to 0$,

$$F(x+h) - F(x) - Vh = O(\|h\|^2).$$

Recall the SVD: $Y = U[\Sigma, 0]V^T = U\Sigma V_1^T$. We have the eigenvalue decomposition

$$\mathcal{S}(Y) := \begin{bmatrix} 0 & Y \\ Y^T & 0 \end{bmatrix} = Q \begin{bmatrix} \Sigma & & \\ & -\Sigma & \\ & & 0 \end{bmatrix} Q^T, \text{ where } Q = \begin{bmatrix} U & U & 0 \\ V_1 & -V_1 & \sqrt{2}V_2 \end{bmatrix}$$

Let $\Pi_+(\cdot)$ be the projector onto the PSD cone, which is known to be strongly semismooth [D.Sun,J.Sun]. Then the strong semismoothness of $D_\rho(\cdot)$ follows from the following result:

$$\begin{aligned} g_\rho(\mathcal{S}(Y)) &= \Pi_+(\mathcal{S}(Y) - \rho I) - \Pi_+(-\mathcal{S}(Y) - \rho I) \\ &= Q \begin{bmatrix} g_\rho(\Sigma) & & \\ & -g_\rho(\Sigma) & \\ & & 0 \end{bmatrix} Q^T = \begin{bmatrix} 0 & D_\rho(Y) \\ D_\rho(Y)^T & 0 \end{bmatrix} = \mathcal{S}(D_\rho(Y)) =: \Psi(Y) \end{aligned}$$

Derivatives of $D_\rho(\cdot)$ (when they exist)

Let Ω the divided difference of $g_\rho(\cdot)$ at the eigenvalue vector λ of $S(Y)$, i.e.,

$$\Omega_{ij} = \frac{g_\rho(\lambda_i) - g_\rho(\lambda_j)}{\lambda_i - \lambda_j}, \quad i,j = 1, \ldots, p+q$$

By [Löwner, 1934], we have

$$\begin{aligned} \Psi'(Y)[H] &= g'_\rho(S(Y))[S(H)] = Q[\Omega \circ (Q^T S(H) Q)]Q^T \qquad (8) \\ &= \begin{bmatrix} 0 & D'_\rho(Y)[H] \\ (D'_\rho(Y)[H])^T & 0 \end{bmatrix} \end{aligned}$$

Let $\alpha = \{1, \ldots, p\}, \quad \gamma = \{p+1, \ldots, 2p\}, \quad \beta = \{2p+1, \ldots, q\}$. By expanding the expression in (8), we get

$$D'_\rho(Y)[H] = U[\Omega_{\alpha\alpha} \circ H_1^s + \Omega_{\alpha\gamma} \circ H_1^a]V_1^T + U(\Omega_{\alpha\beta} \circ H_2)V_2^T$$

where $H_1 = U^T H V_1 = H_1^s + H_1^a$, $H_2 = U^T H V_2$.

PPA. Given a tolerance $\varepsilon > 0$. Input $X^0 \in \mathbb{R}^{p\times q}$ and $\sigma_0 > 0$. Set $k := 0$. Iterate:

Step 1. Compute an approximate maximizer

$$(\zeta^k, \xi^k) \;\approx\; \arg\sup \left\{ \Theta_{\sigma_k}(\zeta, \xi; X^k) \;:\; \zeta \in \mathbb{R}^m,\; \xi \in \mathcal{Q}^* \right\}.$$

Step 2. Compute $W^k := W(\zeta^k, \xi^k; X^k)$. Set

$$X^{k+1} = D_{\rho\sigma_k}(W^k), \quad Z^{k+1} = \frac{1}{\sigma_k}(W^k - D_{\rho\sigma_k}(W^k)).$$

Step 3. If $\|(X^k - X^{k+1})/\sigma_k\| \le \varepsilon$; stop; else; update σ_k; end.

An inexact smoothing Newton method

From now on, we let $\widehat{\mathcal{Q}} := \mathbb{R}^m \times \mathcal{Q} = \mathbb{R}^m \times \mathbb{R}^{s_1} \times \mathbb{R}^{s_2}_+$. Let

$$\widehat{\mathcal{A}} = \begin{bmatrix} \mathcal{A} \\ \mathcal{B} \end{bmatrix}, \; \widehat{b} = \begin{bmatrix} b \\ d \end{bmatrix} \in \Re^{m+s}, \; y = \begin{bmatrix} \zeta \\ \xi \end{bmatrix} \in \widehat{\mathcal{Q}}, \; T = \begin{bmatrix} I_m & 0 \\ 0 & 0_{s\times s} \end{bmatrix}$$

In each **PPA** iteration, for given X and $\sigma > 0$, we need to solve the following subproblem

$$\min_{y\in\widehat{\mathcal{Q}}} \left\{ \theta(y) := \frac{1}{2}\langle y,\, Ty\rangle + \frac{1}{2\sigma}\|D_{\rho\sigma}(W(y;X))\|^2 - \langle \widehat{b},\, y\rangle \right\} \qquad (9)$$

where $W(y;X) = X - \sigma(C - \widehat{\mathcal{A}}^* y)$. We have

$$\nabla\theta(y) = Ty + \widehat{\mathcal{A}} D_{\rho\sigma}(W(y;X)) - \widehat{b}.$$

Since $\theta(\cdot)$ is a convex function, $\bar{y} \in \widehat{\mathcal{Q}}$ solves (9) iff it satisfies the following VI:

$$\langle y - \bar{y},\, \nabla\theta(\bar{y})\rangle \geq 0 \;\forall\, y \in \widehat{\mathcal{Q}} \quad \Leftrightarrow \quad \bar{y} = \Pi_{\widehat{\mathcal{Q}}}(\bar{y} - \nabla\theta(\bar{y})),$$

where $\Pi_{\widehat{\mathcal{Q}}}(\cdot)$ denotes the projector over $\widehat{\mathcal{Q}}$. Define $F : \mathbb{R}^{m+s} \to \mathbb{R}^{m+s}$ by

$$F(y) \;:=\; y - \Pi_{\widehat{\mathcal{Q}}}(y - \nabla\theta(y)) \quad \text{(nonsmooth!)}$$

Then $\bar{y} \in \widehat{\mathcal{Q}}$ solves (9) iff $F(\bar{y}) = 0$.

An inexact smoothing Newton method

Let $h(\varepsilon, t) : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be the Huber smoothing function for $(t)_+ = \max\{t, 0\}$

$$h(\varepsilon, t) = \begin{cases} t & \text{if } t \geq |\varepsilon|/2, \\ \dfrac{1}{2|\varepsilon|}\Big(t + \dfrac{|\varepsilon|}{2}\Big)^2 & \text{if } -|\varepsilon|/2 < t < |\varepsilon|/2, \\ 0 & \text{if } t \leq -|\varepsilon|/2, \end{cases}$$

We use the following smoothing function for $g_\rho(\cdot)$:

$$\boldsymbol{g}_\rho(\varepsilon, t) = h(\varepsilon, t - \rho) - h(\varepsilon, -t - \rho). \tag{10}$$

Then a smoothing function for $D_\rho(Y)$ is

$$\boldsymbol{D}_\rho(\varepsilon, Y) = U\big[\mathrm{Diag}(\boldsymbol{g}_\rho(\varepsilon, \sigma_1), \ldots, \boldsymbol{g}_\rho(\varepsilon, \sigma_p)), \;\; 0\big] V^T,$$

We pick the smoothing function for $\Pi_{\widehat{\mathcal{Q}}}(\cdot)$ to be $\boldsymbol{\pi} : \mathbb{R} \times \mathbb{R}^{m+s} \to \mathbb{R}^{m+s}$: to be

$$\boldsymbol{\pi}_i(\varepsilon, z) = \begin{cases} z_i & \text{if } 1 \leq i \leq m + s_1 \\ h(\varepsilon, z_i) & \text{if } m + s_1 + 1 \leq i \leq m + s \end{cases} \tag{11}$$

Finally, a smoothing function for $F(y) = y - \Pi_{\widehat{\mathcal{Q}}}(y - \nabla\theta(y))$ is given by

$$\boldsymbol{F}(\varepsilon, y) := y - \boldsymbol{\pi}(\varepsilon, y - [\mathcal{T}y + \widehat{\mathcal{A}}\boldsymbol{D}_{\rho\sigma}(\varepsilon, W(y; X)) - \widehat{b}]). \tag{12}$$

We have $F(y) = \boldsymbol{F}(0, y)$ for all y, and $\boldsymbol{F}$ is **strongly semismooth** at $(0, y)$.

An inexact smoothing Newton method

Based on [Gao and Sun 2009] for semidefinite LS problems. Let $\kappa > 0$ be a given constant. Define $E : \mathbb{R} \times \mathbb{R}^{m+s} \to \mathbb{R} \times \mathbb{R}^{m+s}$ by

$$E(\varepsilon, y) := \begin{bmatrix} \varepsilon \\ \overline{\boldsymbol{F}}(\varepsilon, y) := \boldsymbol{F}(\varepsilon, y) + \kappa|\varepsilon| y \end{bmatrix}$$

- $E'(\varepsilon, y)$ is nonsingular for all (ε, y) with $\varepsilon \neq 0$
- E is **strongly semismooth** at $(0, y)$.

Then solving the nonsmooth equation $F(y) = 0$ is equivalent to solving

$$E(\varepsilon, y) = (0, 0).$$

The inexact smoothing Newton method is just Newton-Krylov method applied to minimize the merit function $\|E(\varepsilon, y)\|^2$.

An inexact smoothing Newton method

Step 0. Choose $r \in (0,1)$, $\tau \in (0,1)$, $\hat{\tau} \in [1,\infty)$. Given a starting point (ε^0, y^0), iterate the following steps:

Step 1. Compute

$$\eta := r \min\{1, \|E(\varepsilon^k, y^k)\|^2\}, \quad \hat{\eta} := \min\{\tau, \hat{\tau}\|E(\varepsilon^k, y^k)\|\}.$$

Step 2. Approximately solve the Newton equation $E(\varepsilon^k, y^k) + E'(\varepsilon^k, y^k)[\Delta\varepsilon; \Delta y] = [\eta\varepsilon^0; 0]$ as follows.

Set $\Delta\varepsilon = -\varepsilon^k + \eta\varepsilon^0$.

Apply the BiCGstab method to solve the linear system

$$\boxed{\overline{\boldsymbol{F}}'_y(\varepsilon^k, y^k)\Delta y \;=\; \text{rhs} := -\overline{\boldsymbol{F}}(\varepsilon^k, y^k) - \overline{\boldsymbol{F}}'_\varepsilon(\varepsilon^k, y^k)\Delta\varepsilon}$$

such that the residual R^k satisfies the condition that

$$\|R_k\| \le \min\{\hat{\eta}\|\text{rhs}\|, 0.1\|E(\varepsilon^k, y^k)\|\}$$

Step 3. Apply Armijo linesearch to the merit function $\|E(\varepsilon^k + \alpha\Delta\varepsilon, y^k + \alpha\Delta y)\|^2$ to get a steplength $\bar{\alpha}$.
Set $(\varepsilon^{k+1}, y^{k+1}) = (\varepsilon^k + \bar{\alpha}\Delta\varepsilon, y^k + \bar{\alpha}\Delta y)$.

Quadratic convergence of the inexact smoothing Newton method

- The inexact smoothing Newton method is well defined and generates an infinite sequence $\{(\varepsilon^k, y^k)\}$ such that any accumulation point $(\bar{\varepsilon}, \bar{y})$ is a solution of $E(\varepsilon, y) = 0$ and $\lim_{k\to\infty} \|E(\varepsilon^k, y^k)\| = 0$. Moreover, if Slater's condition holds for NNLS, then $\{(\varepsilon^k, y^k)\}$ is bounded [Gao and Sun 2009].
- To prove the quadratic convergence of $\{(\varepsilon^k, y^k)\}$, it is enough to show that E is strongly semismooth at $(\bar{\varepsilon} = 0, \bar{y})$, and all elements in $\partial_B E(\bar{\varepsilon}, \bar{y})$ are nonsingular.

  Strong semismoothness of E at $(0, \bar{y})$ follows from that of $\boldsymbol{F}$ at $(0, \bar{y})$, and that $|\cdot|$ is strongly semismooth on $\mathbb{R}$.

Constraint nondegeneracy condition for (NNLS)

Let K be the epigraph of $\|X\|_*$, i.e.,

$$K \;:=\; \mathrm{epi}(\|\cdot\|_*) = \{(X;t) \in \mathbb{R}^{p\times q} \times \mathbb{R} \mid \|X\|_* \le t\},$$

which is a closed convex cone. For a given $X_t = (X;t) \in K$, we let $T_K(X_t)$ be the tangent cone of K at X_t, and $\mathrm{lin}(T_K(X_t))$ the largest linear subspace contained in $T_K(X_t)$.

Let $\widehat{\mathcal{B}} := (\mathcal{B},\ 0)$. The problem (NNLS) can be rewritten as:

$$\min\left\{\frac{1}{2}\|\mathcal{A}(X) - b\|^2 + \rho\, t + \langle C, X\rangle \;:\; \widehat{\mathcal{B}}(X;t) \in d + \mathcal{Q},\ (X;t) \in K\right\}. \quad (13)$$

Let $\overline{X}$ be the unique optimal solution to (NNLS). Then $\overline{X}$ is an optimal solution to (13) with $\bar{t} = \|\overline{X}\|_*$. The constraint nondegeneracy condition is said to hold at $(\overline{X};\bar{t})$ if

$$\begin{pmatrix} \widehat{\mathcal{B}} \\ \mathcal{I} \end{pmatrix} (\mathbb{R}^{p\times q} \times \mathbb{R}) + \begin{pmatrix} \mathrm{lin}(T_{\mathcal{Q}}(\widehat{\mathcal{B}}(\overline{X},\bar{t}) - d)) \\ \mathrm{lin}(T_K(\overline{X},\bar{t})) \end{pmatrix} = \begin{pmatrix} \mathbb{R}^s \\ \mathbb{R}^{p\times q} \times \mathbb{R} \end{pmatrix}. \quad (14)$$

Note that $\mathrm{lin}(T_{\mathcal{Q}}(\widehat{\mathcal{B}}(\overline{X},\bar{t}) - d)) = \mathrm{lin}(T_{\mathcal{Q}}(\mathcal{B}(\overline{X}) - d))$.

Let l be the number active inequality constraints at $\overline{X}$. Define $\mathcal{B}^{\text{active}} : \mathbb{R}^{p\times q} \to \mathbb{R}^{s_1+l}$ to be the part of $\mathcal{B}$ corresponding to the active constraints.

Let $W(\bar{y};\overline{X})$ admit the SVD: $U[\Sigma, 0]V^T$. Decompose the index set $\alpha = \{1,\ldots,p\}$ into the following two subsets:

$$\alpha_1 := \{i \mid \sigma_i(W) > \rho\sigma\}, \quad \bar{\alpha}_1 := \alpha\backslash\alpha_1.$$

Then $U = [U_{\alpha_1},\ U_{\bar{\alpha}_1}], V = [V_{\alpha_1},\ V_{\bar{\alpha}_1},\ V_2]$. Consider the following subspace in $\mathbb{R}^{p\times q}$:

$$\mathcal{T}(\overline{X}) := \{H \in \Re^{p\times q} \mid U_{\bar{\alpha}_1}^T\, H\, [V_{\bar{\alpha}_1},\ V_2] = 0\}.$$

Then the constraint nondegeneracy condition (14) can be shown to be equivalent to

$$\mathcal{B}^{\text{active}}(\mathcal{T}(\overline{X})) = \mathbb{R}^{s_1+l}. \tag{15}$$

If the condition (15) holds at $\overline{X}$, then all elements in $\partial_B E(\bar{\varepsilon}, \bar{y})$ are nonsingular.

Some remarks

- When the NNLS problem only has equality constraints, the inner subproblem can be solved by semismooth Newton-CG method.
- The partial PPA (with inexact smoothing Newton) can be applied to semidefinite LS problems with equality/inequality constraints.
- Efficient implementation of partial PPA (with inexact smoothing Newton):
  - Good starting point for partial PPA — we use the alternating direction method of multipliers [Gabay & Mercier 1976, Glowinski & Marrocco 1975] on a reformulation of the NNLS.
  - efficient matrix-vector multiplication for $\boldsymbol{F}'_y(\varepsilon^k, y^k)$
  - preconditioners for the above matrix
  - Implicit computation and storage of V_2, especially when $p \ll q$.

Numerical performance

In our implementation, we apply ADMM to generate a good starting point for the PPA. The stopping criterion for ADMM is $\max\{R_P, R_D\} \leq 10^{-2}$ or that maximal number of 30 iterations is reached.

We stop the PPA when

$$\max\{R_P, R_D\} \leq 10^{-6} \quad \text{and} \quad \text{relgap} := \frac{|\text{pobj} - \text{dobj}||}{1 + |\text{pobj}| + |\text{dobj}|} \leq 10^{-5}$$

Example 1

We consider the approximation problem of $\tilde{M}$ by a low-rank doubly stochastic matrix via solving the following:

$$\min_{X \in \Re^{n \times n}} \left\{ \frac{1}{2} \|X - \widetilde{M}\|^2 + \rho \|X\|_* \ : \ Xe = e,\ X^T e = e,\ X_{11} = M_{11},\ X \geq 0 \right\}.$$

We assume that the observed data is given by $\widetilde{M} = M + \tau N \|M\| / \|N\|$, where τ is the noise factor and N is a random matrix.

For each pair (n, r), we generate a random positive matrix $M \in \mathbb{R}^{n \times n}$ of rank r by setting $M = M_1 M_2^T$ where $M_1 \in \mathbb{R}^{n \times r}$ and $M_2 \in \mathbb{R}^{n \times r}$ have i.i.d. uniform entries in $(0, 1)$. Then M is made doubly stochastic via the Sinkhorn-Knopp algorithm (iteratively perform diagonal scalings on left and right).

Average numerical results over 5 random instances with 10% noise

n / τ	r	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	#sv	time
500 /0.1	10	350148	7.0 | 16.0 | 3.2	1.97e-7 | 1.93e-7 | -6.27e-6	5.42e-2	174	26
	50	501000	5.0 | 9.2 | 2.0	1.65e-7 | 2.31e-7 | -8.58e-6	3.97e-2	177	12
	100	501000	5.0 | 9.0 | 2.1	1.11e-7 | 1.83e-7 | -5.37e-6	3.65e-2	177	12
1000 /0.1	10	1201034	8.0 | 18.8 | 3.6	1.45e-7 | 9.18e-8 | -9.31e-6	5.50e-2	234	2:41
	50	1976915	5.0 | 10.0 | 2.7	7.25e-7 | 7.91e-8 | -3.93e-6	3.30e-2	145	1:13
	100	2002000	3.0 | 6.6 | 2.1	4.43e-7 | 3.32e-7 | -7.58e-6	3.07e-2	143	45
1500 /0.1	10	2552194	9.0 | 22.2 | 3.9	1.69e-7 | 3.84e-8 | -5.68e-6	5.49e-2	275	8:56
	50	3727481	5.0 | 11.0 | 2.7	4.76e-7 | 1.11e-7 | -6.87e-6	3.41e-2	194	3:36
	100	4503000	2.0 | 5.2 | 3.1	2.11e-7 | 2.71e-7 | -3.26e-6	3.19e-2	68	1:55

Example 2

Now consider the low-rank approximation problem of preserving the principal eigenvectors:

$$\min_{X \in \mathbb{R}^{n \times n}} \left\{ \frac{1}{2} \|X - \tilde{M}\|^2 + \rho \|X\|_* \mid Xv = \lambda v, X^T w = \lambda w,\ X \geq 0 \right\}.$$

Average numerical results over 5 random instances with 10% noise

n / τ	r	$m+s$	it.|itsub|bicg	R_p | R_D | relgap	MSE	$r(X)$
500 /0.1	10	350157	2.0 | 5.8 | 2.2	1.85e-7 | 4.88e-7 | -4.58e-6	5.38e-2	170
	50	501000	1.6 | 5.6 | 2.1	4.68e-7 | 8.07e-9 | -4.63e-7	3.94e-2	177
	100	501000	1.8 | 6.2 | 2.1	3.35e-7 | 9.18e-9 | -2.35e-7	3.64e-2	176
1000 /0.1	10	1201029	2.0 | 5.2 | 1.9	6.13e-7 | 2.54e-7 | -2.08e-6	5.28e-2	230
	50	1976912	2.0 | 6.8 | 2.4	9.95e-8 | 1.61e-8 | -5.18e-8	3.27e-2	145
	100	2002000	2.0 | 6.0 | 2.2	9.21e-7 | 1.73e-7 | -2.64e-6	3.04e-2	142
1500 /0.1	10	2552187	2.0 | 5.0 | 1.8	4.56e-7 | 1.83e-7 | 2.16e-6	5.22e-2	278
	50	3727471	2.0 | 5.6 | 2.4	3.95e-7 | 2.93e-8 | 1.75e-7	3.35e-2	192
	100	4503000	2.0 | 7.4 | 2.2	6.33e-8 | 4.31e-8 | -5.30e-7	3.14e-2	67

Given an incomplete, possibly noisy, dissimilarity matrix $B \in \mathcal{S}^n$ with $\mathrm{Diag}(B) = 0$ and sparsity pattern specified by the index set $\mathcal{E}$. The goal is to find an Euclidean distance matrix (EDM) that is nearest to B:

$$\min \Big\{ \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} W_{ij}(D_{ij} - B_{ij})^2 + \frac{\rho}{2n}\langle E,\, D\rangle \;\Big|\; D \text{ is EDM} \Big\},$$

where W_{ij} are given weights, E = matrix of ones.
We added $\frac{\rho}{2n}\langle E,\, D\rangle$ to encourage a sparse solution. From the standard characterization of EDM, we have $D = \mathrm{diag}(X)e^T + e\ \mathrm{diag}(X)^T - 2X$ for some $X \succeq 0$ with $Xe = 0$. The problem can be rewritten as:

$$\min \Big\{ \frac{1}{2} \sum_{(i,j)\in\mathcal{E}} W_{ij}(\langle A_{ij},\, X\rangle - B_{ij})^2 + \rho\langle I,\, X\rangle \;\Big|\; \langle E,\, X\rangle = 0,\ X \succeq 0 \Big\},$$

where $A_{ij} = e_{ij}e_{ij}^T$ with $e_{ij} = e_i - e_j$. Note that desiring sparsity in D leads to the regularization term $\rho\langle I,\, X\rangle$, which is a proxy for desiring a low-rank X.

Regularized kernel estimation (RKE) problem in statistics

We have set of n proteins and dissimilarity measures B_{ij} for certain protein pairs $(i,j) \in \mathcal{E}$ [Lu,Wahba,Wright 05]. The goal is to estimate a positive semidefinite kernel matrix $X \in \mathcal{S}^n_+$ such that the fitted squared distances induced by X for the protein pairs satisfy

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

problem	n	m	ρ	it.	itsub	cg	R_p	R_D	relgap	#sv	ti
RKE630	630	198136	5.07e-1	6	36	24.6	1.07e-7	2.42e-8	-1.81e-6	388	1:
PDB25	1898	1646031	1.84e+0	18	55	55.8	4.89e-7	4.78e-6	-1.46e-5	1388	1:19

Conclusion & Future Work

- We introduced a proximal point algorithm for solving nuclear norm regularized matrix LS problems with a large number of equality and inequality constraints
- The inner subproblems are solved by an inexact smoothing Newton method, which is proved to be quadratically convergent under the constraint nondegeneracy condition.
- Numerical experiments on selected examples demonstrated that our PPA based algorithm is efficient.
- Our framework can be extended to LS problems with other regularizers such as $\|X\|_2$, cone of epi-graph of "nice" norm, mixed-norm like $\sum_{k=1}^{N} \|X_k\|_2$, etc. (As long as the associated proximal-point operator can be computed efficiently).

Thank you for your attention!