Orthogonal constrained optimization problem based on augmented Lagrangian method

Jinchao Zhang

Hebei University of Technology

2021.1.27

Contents

1. W. Chen, H. Ji, and Y. You, An augmented lagrangian method for ℓ_1 regularized optimization problems with orthogonality constraints[J], SIAM J. Sci. Comput., 38 (2016), pp. B570–B592.
2. Bin Gao, Xin Liu, Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints[J]. SIAM Journal on Scientific Computing, 41-3 (2019), A1949 – A1983.
3. Nachuan Xiao, Xin Liu, Ya-xiang Yuan. Exact penalty function for $\ell_{2,1}$ norm minimization over the stiefel manifold.2020
4. Stage work summary.

Augmented lagrangian method for orthogonality constraints optimization problems

Consider the following problem:

$$\begin{aligned} &\min_{X\in\mathbb{R}^{n\times p}} F(X) = f(X) + g(X) \\ &s.t \quad X^T X = I_p \end{aligned} \tag{1}$$

Assumption:

(1) F might be nonconvex and nondifferentiable;

(2) f is possibly nonconvex, gradient ∇f is Lipschitz continuous,g is convex ,possibly nonsmooth .

Standard augmented lagrangian method yields the following scheme:

$$\begin{cases} X^{k+1} \in \arg\min F(X) + \frac{\rho^k}{2} \left\| X^\top X - I_p \right\|_F^2 + \mathrm{Tr}\left(\left(\Lambda^k\right)^\top \left(X^\top X - I_p \right) \right) \\ \Lambda^{k+1} = \Lambda^k + \rho^k \left(\left(X^{k+1}\right)^\top X^{k+1} - I_p \right) \end{cases} \tag{2}$$

- The subproblem of the above augmented Lagrangian scheme is rather complex;
- generally has no analytic solution.

PAMAL method

Rewrite the optimization problem:

$$\min_{X,Q,P\in\mathbb{R}^{n\times m}} f(X)+g(Q)+\delta_{S_{n,p}}(P) \tag{3}$$

$$\text{s.t. } Q-X=0 \quad \text{and} \quad P-X=0 \tag{4}$$

where $\delta_{S_{n,p}}$ is indicator function defined by

$$\delta_{S_{n,p}}(X)=\begin{cases} 0 & \text{if } X\in S_{n,p} \\ +\infty & \text{otherwise} \end{cases} \tag{5}$$

Augmented Lagrangian function:

$$L(X,Q,P,\Lambda;\rho) = f(X)+g(Q)+\langle\Lambda_1, Q-X\rangle+\frac{\rho}{2}\|Q-X\|_F^2 \tag{6}$$

$$+\langle\Lambda_2, P-X\rangle+\frac{\rho}{2}\|P-X\|_F^2+\delta_{\mathcal{S}}(P) \tag{7}$$

where ρ is positive penalty paramter,$\Lambda:=(\Lambda_1,\Lambda_2)\subset\mathbb{R}^{n\times 2m}$.

PAMAL method

Algorithm 1 (Method for solving (6))

Input: Given predefined parameters $\{\epsilon^k\}_{k\in\mathbb{N}}$,$\bar{\Lambda}^1 := (\bar{\Lambda}_1^1, \bar{\Lambda}_2^1)$,$\rho^1$, $\bar{\Lambda}_{i,min}$, $\bar{\Lambda}_{i,max}$,τ, γ that satisfy the certain conditions,for $k = 1, 2...$,

Step 1: Compute (X^k, Q^k, P^k) such that there exists $\Theta^k \in \partial L(X^k, Q^k, P^k, \bar{\Lambda}^k)$; ρ^k satisfying

$$\|\Theta^k\|_\infty \leq \epsilon^k, (P^k)^T P^k = I_p;$$

Step 2: Update the multiplier estimates:

$$\Lambda_1^{k+1} = \bar{\Lambda}_1^k + \rho^k(Q^k - X^k), \qquad \Lambda_2^{k+1} = \bar{\Lambda}_2^k + \rho^k(P^k - X^k)$$

where $\bar{\Lambda}_i^{k+1}$ is the projection of Λ_i^{k+1} on $\{\Lambda_i : \bar{\Lambda}_{i,min} \leq \Lambda_i \leq \bar{\Lambda}_{p,max}\}, i = 1, 2$

Step 3: Update the penalty parameter:

$$\rho^{k+1} := \begin{cases} \rho^k & \text{if } \left\|R_i^k\right\|_\infty \leq \tau \left\|R_i^{k-1}\right\|_\infty, i = 1, 2 \\ \gamma\rho^k & \text{otherwise} \end{cases}$$

where $R_1^k := Q^k - X^k$, $R_2^k := P^k - X^k$

PAM method for Step 1

For kth outer iteration,the inner iterations can be viewed as a proximal regularization of a there-block Gauss-Seidel method:

$$\begin{cases} X^{k,j} \in \arg\min_X L\left(X, Q^{k,j-1}, P^{k,j-1}, \bar{\Lambda}^k; \rho^k\right) + \frac{c_1^{k,j-1}}{2}\left\|X - X^{k,j-1}\right\|_F^2 \\ Q^{k,j} \in \arg\min_Q L\left(X^{k,j}, Q, P^{k,j-1}, \bar{\Lambda}^k; \rho^k\right) + \frac{c_2^{k,j-1}}{2}\left\|Q - Q^{k,j-1}\right\|_F^2 \\ P^{k,j} \in \arg\min_P L\left(X^{k,j}, Q^{k,j}, P, \bar{\Lambda}^k; \rho^k\right) + \frac{c_3^{k,j-1}}{2}\left\|P - P^{k,j-1}\right\|_F^2, \end{cases} \tag{8}$$

where proximal parameters $\{c_i^{k,j}\}_{k,j}$,satisfy $0 < \underline{c} \le c_i^{k,j} \le \bar{c} < \infty, i = 1, 2, 3$
Termination conditions: $\Theta^{k,j} \in \partial L(X^{k,j}, Q^{k,j}, P^{k,j}, \bar{\Lambda}^k; \rho^k)$ satisfying

$$\|\Theta^{k,j}\|_\infty \le \epsilon^k, \quad (P^{k,j})^T P^{k,j} = I_p$$

[1] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the Kurdyka-Lojasiewicz inequality. Mathematics of Operations Research, 35(2):438 – 457, 2010.

PAM method for Step 1

$\Theta^{k,j} := (\Theta_1^{k,j}, \Theta_2^{k,j}, \Theta_3^{k,j})$ is given by

$$\begin{cases} \Theta_1^{k,j} := \rho^k \left(Q^{k,j-1} - Q^{k,j} + P^{k,j-1} - P^{k,j} \right) + c_1^{k,j-1} \left(X^{k,j-1} - X^{k,j} \right) \\ \Theta_2^{k,j} := c_2^{k,j-1} \left(Q^{k,j-1} - Q^{k,j} \right) \\ \Theta_3^{k,j} := c_3^{k,j-1} \left(P^{k,j-1} - P^{k,j} \right) \end{cases} \tag{9}$$

PAM method for Step 1

Algorithm 2 (Method for solving **Algorithm 1's step 1**)

Input: Let $(X^{1,0}, Q^{1,0}, P^{1,0})$ be any initialization. For $k \geq 2$, set

$$(X^{k,0}, Q^{k,0}, P^{k,0}) := (X^{k-1}, Q^{k-1}, P^{k-1})$$

Step 1: Compute (X^k, Q^k, P^k) such that there exists $\Theta^k \in \partial L(X^k, Q^k, P^k, \bar{\Lambda}^k); \rho^k$ satisfying

$$\|\Theta^k\|_\infty \leq \epsilon^k, (P^k)^T P^k = I_p;$$

Step 2: Reiterate on j until $\|\Theta^{k,j}\|_\infty \leq \epsilon^k$

1. $X^{k,j} = Z^{-1}\left(\bar{\Lambda}_1^k + \bar{\Lambda}_2^k + \rho^k Q^{k,j-1} + \rho^k P^{k,j-1} + c_1^{k,j-1} X^{k,j-1}\right)$, where
$$Z := Z^{k,j-1} = 2H + \left(\rho^k + \rho^k + c_1^{k,j-1}\right) I_n$$
2. $Q^{k,j} = T_\eta^1\left(\frac{\rho^k X^{k,j} - \bar{\Lambda}_1^k + c_2^{k,j-1} Q^{k,j-1}}{\rho^k + c_2^{k,j-1}}\right), \eta := \eta^{k,j} := \mu \cdot \left(\rho^k + c_2^{k,j-1}\right)^{-1}$, where T_η^1 is the soft-thresholding operator.
3. $P^{k,j} = U I_{n\times m} V^\top$, where the matrices U, V are obtained from the SVD of
$$\frac{\rho^k X^{k,j} + c_3^{k,j-1} P^{k,j-1} - \bar{\Lambda}_2^k}{\rho^k + c_3^{k,j-1}} =: U\Sigma V^\top$$

Step 3: Set $(X^k, Q^k, P^k) := (X^{k,j}, Q^{k,j}, P^{k,j})$

$$\Theta^k := \Theta^{k,j}$$

Convergence analysis

Theorem 1

[2,Th 6.2] Suppose that F is a K-L function of the original problem. Let $\{X^k\}_{k\in\mathbb{N}}$ be a sequence generated by [3,Algorithm 4].If the sequence $\{X^k\}_{k\in\mathbb{N}}$ is bounded,then the folllowing assertions hold:

(1) The sequence $\{X^k\}_{k\in\mathbb{N}}$ has finite length;i.e., $\sum_{k=1}^{\infty}\|x^{k+1}-x^k\|_F<\infty$

(2) The sequence $\{X^k\}_{k\in\mathbb{N}}$ converges to a critical point $\bar{X}$ of F.

The global convergence of the PAM method established requires the objective function F to satisfy the K-L property.

[2] H. Attouch, J. Bolte, and B. F. Svaiter, Convergence of descent methods for semi-algebraic and tame problems: Proximal algorithms, forward backward splitting, and regularized Gauss-Seidel methods, Math. Program., 137 (2013), pp.91-129.

Convergence analysis

Theorem 2

[3,Th 3]Let $\sigma : \mathbb{R}^d \to (-\infty, +\infty]$ *be a proper and lower semiccontinuous function.*

If σ *is semi-algebraic then it satisfies the KL proper at any point of dom*σ

Theorem 3

Suppose that the positive parameters γ, ρ^1 *in Algorithm 1 are chosen so that* $\gamma > 1, 2H + \rho^1 I_n \succ 0$. *Let* $\{(X^k, Q^k, P^k)\}_{k\in\mathbb{N}}$ *be the sequence generated by Algorithm 1. Then the sequence* $\{(X^k, Q^k, P^k)\}_{k\in\mathbb{N}}$ *is nonempty,and every limit point is KKT point of the original problem.*

[3]J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program., 146 (2014), pp. 459–494.

Numerical experiments

Compressed modes problem:

$$\min_{X,Q,P} \mu\|Q\|_1 + \mathrm{Tr}\left(X^T H X\right)$$
$$\text{s.t. } Q - X = 0, P - X = 0, P^\top P = I$$

Problems			No. of iterations		CPU time (s)	
n	p	μ	PAMAL	SOC	PAMAL	SOC
128	8	0.5	433	848	0.28	0.41
128	10	0.5	833	2215	0.60	1.33
64	10	0.3	697	1398	0.17	0.24
128	10	0.3	921	2393	0.67	1.67

[4] R. Lai and S. Osher, A splitting method for orthogonality constrained problems[J], J. Sci. Comput., 58 (2014), pp. 431–449.

The second thesis

Consider the following problem:

$$\begin{aligned} &\min_{X \in \mathbb{R}^{n \times p}} f(X) \\ &s.t \quad X^T X = I_p \end{aligned} \tag{10}$$

where $f : \mathbb{R}^{n \times p} \mapsto \mathbb{R}$ is a continuously differentiable function.

Main motivation and contribution:

- Orthogonalization processes are expensive to compute and difficult to parallelize;
- Based on the augmented Lagrangian method (ALM),the dual variables enjoy a closed-form formula at each first-order stationary point;
- For the prime variables,minimize a proximal linearized approximation of the augmented Lagrangian function, which is equivalent to taking one gradient descent step;

[5]Bin Gao, Xin Liu, Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints[J]. SIAM Journal on Scientific Computing, 41-3 (2019), A1949 – A1983.

Augmented Lagrangian method

Algorithm 3 (Augmented Lagrangian method (ALM))

Input:choose initial guess Λ^0 for the dual variables, and set $k := 0$;

while certain stopping criterion is not reached **do**

Step 1: Minimize the augmented Lagrangian function

$$X^{k+1} := \min_{X \in \mathbb{R}^{n \times p}} \quad \mathcal{L}_\beta(X, \Lambda^k);$$

where the augmented Lagrangian function is defined as

$$\mathcal{L}_\beta(X, \Lambda) = f(X) - \frac{1}{2}\left\langle \Lambda, X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathrm{F}}^2$$

Step 2: Update the Lagrangian multipliers :

$$\Lambda^{k+1} := \Lambda^k - \beta((X^{k+1})^T X^{k+1} - I_p)$$

Step 3: Update the penalty parameter β if necessary.Set $k := k + 1$

Output: X^k

The optimality condition

The Augmented Lagrangian function of (12)

$$L_{X,\Lambda} := f(X) - \tfrac{1}{2}\langle \Lambda, X^T X - I \rangle \tag{11}$$

where $\Lambda \in \mathbb{S}^{p\times p}$ can be viewed as the Lagrangian multipliers.

$$\nabla_X L(X,\Lambda) = \nabla f(X) - X\Lambda = 0 \Longrightarrow \Lambda = X^T \nabla f(X)$$

Definition 1 (first order optimality condition)

Given a point $X \in \mathbb{R}^{n\times p}$,

$$\begin{cases} \left(I_n - XX^\top\right)\nabla f(X) = 0 & \textbf{\textit{substationarity,}} \\ X^\top \nabla f(X) = \nabla f(X)^\top X & \textbf{\textit{symmetry,}} \\ X^\top X = I_p & \textbf{\textit{feasibility.}} \end{cases} \tag{12}$$

Updating multipliers by closed form:

$$\Lambda^{k+1} := \Phi(\nabla f(X^k)^T X^k)$$

where $\Phi : R^{n\times n} \mapsto \mathbb{S}^n$ is defined $\Psi(A) := \frac{1}{2}(A + A^T)$

The proximal linearized augmented Lagrangian algorithm

Algorithm 4 (Proximal linearized augmented Lagrangian algorithm (PLAM))

Input:choose initial guess X^0 for the dual variables, and set $k := 0$;

while certain stopping criterion is not reached **do**

Step 1: Compute the Lagrangian multipliers

$$\Lambda^k := \Psi(\nabla f(X^k)^T X^k).;$$

Step 2: Minimize the following proximal linearized Lagrangian function

$$X^{k+1} := \underset{X \in \mathbb{R}^{n\times p}}{\arg\min} \tilde{\mathcal{L}}_\beta(X) = \mathbf{tr}\left(\nabla_X \mathcal{L}_\beta\left(X^k, \Lambda^k\right)^\top \left(X - X^k\right)\right) + \frac{\eta^k}{2}\left\|X - X^k\right\|_{\mathrm{F}}^2$$

Step 3: Set $k := k + 1$

Output:: X^k

where **Step 2** actually is gradient step

$$X^{k+1} = X^k - \frac{1}{\eta^k}\nabla_X \mathcal{L}_\beta\left(X^k, \Lambda^k\right) \tag{13}$$

[6]J. Bolte, S. Sabach, and M. Teboulle, Proximal alternating linearized minimization for nonconvex and nonsmooth problems, Math. Program., 146 (2014), pp. 459–494.

Lemma 1

For any X^* satisfying $\sigma_{min}(X^*) > 0$, suppose,

$$\beta > (\|\nabla f(X^*)\|_2 \cdot \|X^*\|_2 + \delta)/\sigma_{min}^2(X^*)$$

with $\delta > 0$. Then it holds that

$$\left\|X^{*\top} X^* - I_p\right\|_{\mathbf{F}} \leq \frac{\|X^*\|_2}{\delta} \cdot \|\nabla_X \mathcal{L}_\beta\left(X^*, \Lambda^*\right)\|_{\mathbf{F}} \tag{14}$$

with $\Lambda^* = \Psi(\nabla f(X^*)^T X^*)$. In particular,if it happens that X^* is a first-order sationary point of

$$\min_{X \in \mathbb{R}^{n \times p}} L_\beta(X, \Lambda^*)$$

Then X^* is also a first-order stationary point of original problem

where $\Psi(X) = \frac{X + X^T}{2}$

Convergence of PLAM

Special notations:

$$R = \left\|X^{0\top}X^0 - I_p\right\|_{\mathrm{F}}; \quad \mathcal{C} = \left\{X \mid \left\|X^\top X - I_p\right\|_{\mathrm{F}} \leq R\right\}; \quad \underline{f} = \min_{X\in\mathcal{C}} f(X)$$
$$M = \max_{X\in\mathcal{C}} \|X\|_2; \quad N = \max_{X\in\mathcal{C}} \|\nabla f(X)\|_{\mathrm{F}}; \quad L = \max_{X\in\mathcal{C}} \left\|\nabla^2 f(X)\right\|_2$$

Merit function:

$$h(X) = f(X) - \frac{1}{2}\left\langle \Psi\left(\nabla f(X)^\top X\right), X^\top X - I_p\right\rangle + \frac{\beta}{4}\left\|X^\top X - I_p\right\|_{\mathrm{F}}^2 \tag{15}$$

- f is twice continuous differentiable;
- ∇f is Lipschitz continuous on the compact set $\mathcal{C}$,there exits constant $L_h > 0$,related to β;

$$\|\nabla h(X) - \nabla h(y)\|_{\mathrm{F}} \leq L_h \|X - Y\|_{\mathrm{F}} \quad \forall X, Y \in \mathcal{C} \tag{16}$$

Global convergence of PLAM

Theorem 4 (the worst case complexity)

Suppose $\{X^k\}$ *is the iterate sequence generated by Algorithm PLAM initiated from* X^0 *and the problem parameters satisfy Assumption. Then the sequence* $\{X^k\}$ *has at least one cluster point, and any cluster point is a first-order stationary point of original problem. More precisely, for any* $K > 1$*,it holds that*

$$\min_{k=0,\ldots,K-1} \left\|\nabla_X \mathcal{L}_\beta\left(X^k, \Lambda^k\right)\right\|_{\mathrm{F}} < \sqrt{\frac{f\left(X^0\right) - \underline{f} + MNR + \beta R^2/4}{c_3 K}} \tag{17}$$

★ Sublinear convergence rate;

★ Algorithm terminates after $O(1/\epsilon^2)$ iterations if the stopping criterion is set as

$$\max\left\{\left\|\nabla_X \mathcal{L}_\beta\left(X^k, \Lambda^k\right)\right\|_{\mathrm{F}}, \left\|{X^k}^\top X^k - I\right\|_{\mathrm{F}}\right\} < \epsilon \tag{18}$$

Local convergence rate of PLAM

Theorem 5

Suppose $\{X^*\}$ *is an isolated minimizer of original problem,and we denote*

$$\tau := \inf_{0 \neq Y \in \mathcal{T}_{\mathcal{X}}(S_{n,p})} \frac{\mathrm{tr}\left(Y^\top \nabla^2 f(X)[Y] - \Lambda Y^\top Y\right)}{\|Y\|_{\mathrm{F}}^2} \tag{19}$$

The algorithm parameters satisfy $\beta \geq \frac{L+MN+\tau}{2}$ *and* $\eta^k \in [\underline{\eta}, \bar{\eta}]$,*where* $\bar{\eta} \geq \underline{\eta} \geq L + MN + 2\beta$. *Then,there exists* $\epsilon > 0$ *such that starting from any* X^0 *satisfying* $\|X^0 - X^*\| < \epsilon$,*the iterate sequence* $\{X^k\}$ *generated by PLAM converges to* X^* *Q-linearly.*

Upgraded version of PLAM.

The limitations of PLAM

- The numerical performance is very sensitive to parameters β_k and η_k;
- If β_k smaller $\longrightarrow$ No convergence;
- If β_k be sufficiently large then need lager η_k $\longrightarrow$ will Slow convergence.

★ Strategy: impose redundant columnwise unit sphere constraints.

$$\begin{aligned} \min_{X\in\mathbb{R}^{n\times p}} \tilde{\mathcal{L}}_\beta(X) &:= \left\langle \nabla_X \mathcal{L}_\beta\left(X^k,\Lambda^k\right), X-X^k\right\rangle + \frac{\eta_k}{2}\left\|X-X^k\right\|_{\mathrm{F}}^2 \\ \text{s. t. } & \|X_i\| = 1, \quad i=1,\ldots,p \end{aligned} \tag{20}$$

Corresponding Lagrange multiplier update:

$$\Lambda^k := \Psi\left(\nabla f\left(X^k\right)^\top X^k\right) + \Phi\left(X^{k\top}\nabla_X L_\beta\left(X^k, \Psi\left(\nabla f\left(X^k\right)^\top X^k\right)\right)\right) \tag{21}$$

where $\Phi(M) := Diag(diag(M))$

Parallelizable Column-wise Block Minimization for PLAM (PCAL).

Algorithm 5 (PCAL)

Input: choose initial guess X^0 for the dual variables, and set $k := 0$;

while certain stopping criterion is not reached **do**

Step 1: Compute the Lagrangian multipliers

$$\Lambda^k := \Psi(\nabla f(X^k)^T X^k) + \Phi(X^{k^T} \nabla_X L_\beta(X^k, \Psi(\nabla f(X^k)^T X^k)));$$

Step 2: **for** $i = 1, ..., P$ **do**

$$X_i^{k+1} := \frac{X_i^k - \frac{1}{\eta_k} \nabla_{X_i} \mathcal{L}_\beta\left(X^k, \Lambda^k\right)}{\left\| X_i^k - \frac{1}{\eta_k} \nabla_{X_i} \mathcal{L}_\beta\left(X^k, \Lambda^k\right) \right\|_2}$$

then $X^{k+1} := [X_1^{k+1}, ..., X_p^{k+1}]$;

Step 3: Set $k := k + 1$

Output: X^k

The third thesis

General form:

$$\begin{aligned} &\min_{X\in\mathbb{R}^{n\times p}} F(X) = f(X) + r(X) \\ &s.t \quad X^T X = I_p \end{aligned} \tag{22}$$

- $f : \mathbb{R}^{n\times p} \longmapsto \mathbb{R}$;
- $r(X) = \sum_{i=1}^{n} \gamma_i \|X(i,:)\|_2$;
- Stiefel manifold:

$$S_{n,p} := \{X \in \mathbb{R}^{n\times p} | X^T X = I_p\}$$

[7]N. Xiao, X. Liu and Y. Yuan, Exact penalty function for $\ell_{2,1}$ norm minimization over the stiefel manifold, SIAM Journal on Optimization, optimization online: 2020/07/7908.

Motivation

The latest approach:proximal gradient approaches

- Proximal gradient method on manifold (ManPG):Chen-Ma-So-Zhang 2020;

$$\min_{D \in T_{St_{n,p}}(X)} \langle D, \nabla f(X_k) \rangle + r(X_k + D) + \frac{\|D\|_F^2}{2\eta_k} \quad (\text{ proximal mapping })$$

Computational cost per outer iteration

- No closed-form solution for proximal mapping $\Rightarrow$ semismooth Newton method;
- Orthonormalization process is required in each iteration $\Rightarrow$ lacks scalability;

[8]Chen, Shixiang, et al. "Proximal gradient method for nonsmooth optimization over the stiefel manifold[J]." Siam Journal on Optimization, vol. 30, no. 1, 2020, pp. 210 – 239.

Penalty Function

Exact penalty function

$$h(X) := f(X) + r(X) - \frac{1}{2}\left\langle \Lambda(X), X^\top X - I_p \right\rangle + \frac{\beta}{4}\left\| X^\top X - I_p \right\|_{\mathbf{F}}^2 \qquad (23)$$

where

$$\Lambda(X) = \Phi\left(X^\top \nabla f(X)\right) + \sum_{i=1}^{n} \gamma_i S\left(X_i\right)$$

X^* is a first-order stationary point $\Rightarrow 0 \in \partial h(X^*)$;

$$0 \in \partial h(X) \Rightarrow X^T X = I_p?$$

$h(X)$ may be unbounded from below - the example in smooth case is also true for nonsmooth case

- $f(X) = \frac{1}{4}\|X^T X - I_p\|_F^2, \gamma_i = 0$;
- $h(X) = \frac{1}{4}\|X^T X\|_F^2 - 2tr((X^T X)^2(X^2 X - I_p)) + \frac{\beta}{4}\|X^T X - I_p\|_F^2$
- $\|X\|_F \to +\infty \Rightarrow h(X) \to -\infty$.

Properties of Penalty Model

Constants

- $M_0 := \sup_{X \in \mathcal{M}} \|\nabla f(X)\|_{\mathbf{F}}$;
- $L_1 := \sup_{X \in \mathcal{M}, Y \in \mathcal{M}} \frac{\|\Lambda(X) - \Lambda(Y)\|_F}{\|X - Y\|_{\mathbf{F}}}$
- $M_1 := \sup_{X \in \mathcal{M}} \|\Lambda(X)\|_2$;
- $C_1 := \sup_{X \in \mathcal{M}} \tilde{h}(X) - \inf_{X \in \mathcal{M}} \tilde{h}(X)$

where $\tilde{h}(X) = f(X) + r(X) - \frac{1}{2}\langle \Lambda(X), X^T X - I_p \rangle$

Lemma 2

For any $0 < \delta \leq \frac{1}{3}$, *when* $\beta \geq \max\{2(M_0 + M_1), 2pL_1, (3M_1 + \frac{3\sqrt{2}}{2}L_1), \frac{2C_1}{\delta^2}\}$, *we have*

$$\sup_{\|X^\top X - I_p\|_{\mathbf{F}} \leq \delta} h(X) < \inf_{\|X^\top X - I_p\|_{\mathbf{F}} \geq 2\delta} h(X) \tag{24}$$

Moreover any global minimizer X^* *of(PenC) satisfies* $X^* \in S_{n,p}$, *which further implies that it is a global minimizer of original problem.*

A New Penalty Model

Restrict h in a bounded set

$$\min_{X \in \mathcal{M}} h(X) \qquad (PenC) \tag{25}$$

- $\mathcal{M}$ is a convex compact set, $S_{n,p} \subset \mathcal{M}$
- Ball with radius K in F-norm: $\mathcal{B} := \{X \in \mathbb{R}^{n \times p} | \|X\|_F \leq K\}$

Assumption 1

f(X) is differentiable and $\nabla f(X)$ is Lipschitz continuous.

Problem reformulation

$$\begin{array}{ll} \min_{X \in \mathbb{R}^{n \times p}} & f(X) + r(X) \\ \text{s.t.} & X^\top X = I_p, \end{array} \Longrightarrow \min_{X \in \mathcal{B}} h(X) \tag{26}$$

$\mathcal{B} := \{X \in \mathbb{R}^{n \times p} | \|X\|_F \leq K\}$ where $K > \sqrt{p}$.

First-order Optimal Condition

Definition 2

A point $X \in S_{n,p}$ is called as first order stationary point if and only if it satisfies

$$0 \in grad(\nabla f(X) + \partial r(X)) = \mathcal{P}_{\mathcal{T}_X}(\nabla f(X) + \partial r(X))$$

where $\mathcal{T}_X$ denotes the tangent space at X, $\mathcal{P}_{\mathcal{T}}(\cdot)$ consists of all the projection points of $(\cdot)$ onto the tangent space $\mathcal{T}_X$.

Equivalent version

- There exists $D \in \partial r(X)$ and $\Lambda \in \mathbb{R}^{p\times p}$:

$$\begin{cases} X\Lambda = \nabla f(X) + D \\ \Lambda = \Lambda^\top \\ X^\top X = I_p \end{cases} \tag{27}$$

$$\Lambda(X) \in \Phi(X^T \nabla f(X) + X^T \partial r(X))$$

Lagrange multiplier explicit expression

Expression for $\partial r(X)$

- $\partial r(X) = [\gamma_1 \partial \left(\|X_1 \cdot\|_2\right), \gamma_2 \partial \left(\|X_2 \cdot\|_2\right), \ldots, \gamma_n \partial \left(\|X_n \cdot\|_2\right)]^\top$;
- $\partial \left(\|X_j \cdot\|_2\right) = \begin{cases} \frac{X_j^\top}{\|X_j \cdot\|_2}, & \text{if } \|X_j\|_2 \neq 0 \\ u_j \text{ satisfying } \|u_j\|_2 = 1, & \text{otherwise} \end{cases}$
- For any $D \in \partial r(X)$,

$$X^\top D = \sum_{i=1}^{n} \gamma_i S\left(X_i\right), \quad \text{where } S(x) := \begin{cases} \frac{xx^\top}{|x|_2}, & \text{if } x \neq 0 \\ 0, & \text{otherwise} \end{cases}$$

- $S(X)$ Lipschitz continuous. Hence,

$$\Lambda(X) = \Phi\left(X^\top \nabla f(X)\right) + \sum_{i=1}^{n} \gamma_i S\left(X_i\right)$$

- Closed-form expression;
- Lipschitz continuous.

Algorithm

Difficulties in computing proximal mapping

$$h(Y) := f(Y) + \frac{\beta}{4}\|Y^T Y - I_p\|_F^2 - \frac{1}{2}\langle \Lambda(Y), Y^T Y - I_p\rangle + r(Y) \tag{28}$$

- $f(Y) + \frac{\beta}{4}\|Y^T Y - I_p\|_F^2$:smooth
- $r(Y)$:nonsmooth
- $\frac{1}{2}\langle \Lambda(Y), Y^T Y - I_p\rangle$:nonconvex,nonsmooth $\Rightarrow$ approximate

$$\langle Y - X^k, X^k \Lambda(X^k)\rangle + \tfrac{1}{2}\langle \Lambda(X^k), {X^k}^T X^k - I_p\rangle \tag{29}$$

- When feasibility violation $\|Y^T Y - I_p\|$ is same order of $\|Y - X\|_F$,second order approximation.

$$D_k = \nabla f(X_k) + \beta X_k[({X^k}^T X_k - I_p) - \Lambda(X_k)] \tag{30}$$

- using the approximate (29,30),

$$Y_k = \arg\min_{Y\in\mathbb{R}^{n,p}} \langle Y - X_k, D_k\rangle + r(Y) + \frac{\|Y - X_k\|_F^2}{2\eta_k}$$

Proximal Gradient Method for Solving PenC with Exact Lambda (PenCPG)

Algorithm 8 (Proximal Gradient Method for PenC (PenCPG))

Input: choose initial guess X_0 , $\beta > 0$ and set $k := 0$;

while certain stopping criterion is not reached **do**

Step 1: Compute the direction D_k

Step 2: Choose stepsize η^k by certain strategy;

Step 3: Update Y_k by

$$Y_k = \arg\min_{Y \in \mathbb{R}^{n,p}} \langle Y - X_k, D_k \rangle + r(Y) + \frac{\|Y - X_k\|_F^2}{2\eta_k};$$

Step 4: **if** $\|Y_k\|_F > K$, **then**

$$X_{k+1} = \frac{K}{\|Y_k\|_F} Y_k;$$

else

$$X_{k+1} = Y_k;$$

end if

end while

Output: X^k

Global Convergence

Theorem 6

Suppose Assumption 3 holds.Let $0 < \delta \leq \frac{1}{3}, K \geq \frac{\sqrt{6p}}{2}$ *and* $\beta \geq \max\{6M_1, \max\{2p, 12\sqrt{6}\}L_1, 2(M_0 + M_1, \frac{2C_1}{\delta^2})\}$ *Suppose that* $\{X_k\}$ *is the iterate sequence generated by PenCPG,starting from the initial point* $X_0 \in \mathcal{B}$ *satisfying* $\|X_0^T X_0 - I_p\|_F \leq \frac{\delta}{2}$,*and adopting the stepsize* $\eta_k \in [\frac{1}{2}\eta^+, \eta^+]$ *where*

$$\eta^+ = \min\left\{\frac{1}{L_0 + 4\beta + \frac{\sqrt{6}}{2}L_1 + M_1}, \frac{1}{15\left(M_0 + \frac{2\sqrt{3p}}{3}M_1 + \frac{2\sqrt{3}}{9}\beta + L_r\right)}, \frac{1}{4\left(L_0 + 4\beta + M_1\right)}\right\}$$

Then $\{X_k\}$ *exists clustering point and any clustering point is a first-order stationary point of original problem. More precisely,for any* $N \geq 1$,*it holds that*

$$\min_{0 \leq k \leq N-1} \|X_{k+1} - X_k\|_F \leq \sqrt{\frac{(16C_1 + \beta\delta^2)\eta^+}{2N}}$$

The subliner convergence rate of $\|X_{k+1} - X_k\|_F$ illustrated in the above theorem actually tells us that PenCPG terminates after $O(\frac{1}{\epsilon^2})$

Numerical experiments

Sparse variable PCA

$$\min_{X \in \mathbb{R}^{n \times p}} -\frac{1}{2}\operatorname{tr}\left(X^\top M X\right) + \|\Gamma X\|_{2,1} \quad \text{s.t.} \quad X^\top X = I_p \tag{31}$$

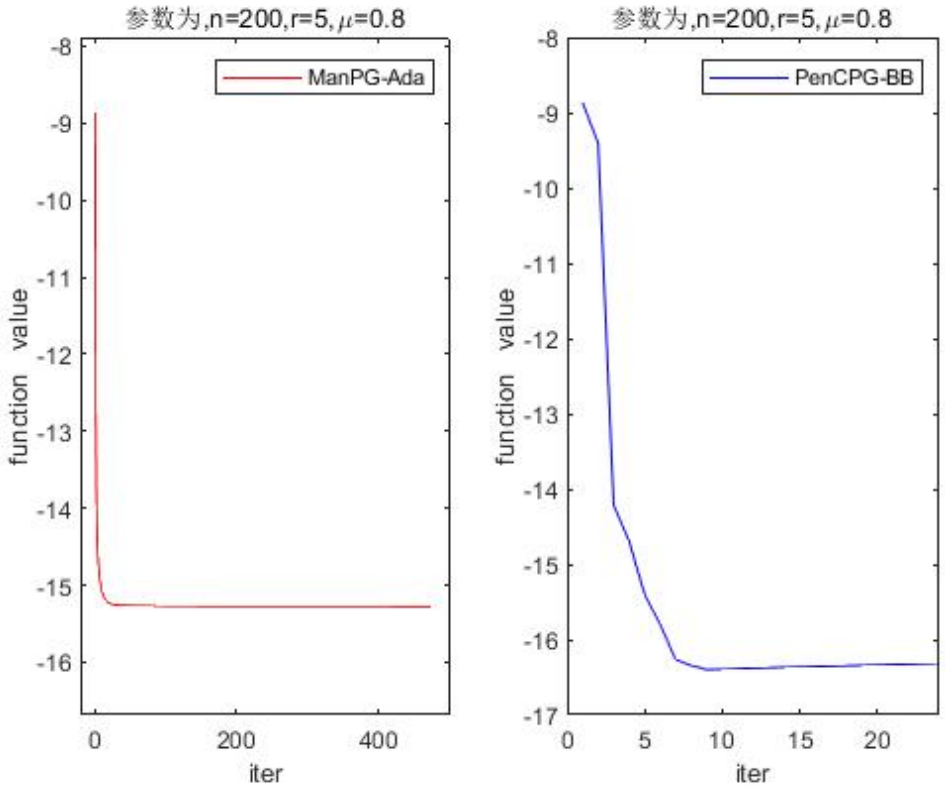


Figure 1: Comparison function value of ManPG-Ada ,PenCPG-BB

Stage work summary

Algorithm 9 (Manifold proximal gradient method (ManPG))

Input:initial point $X_0 \in M$,$\gamma \in (0,1)$,stepsize $t > 0$.
for $k = 0, 1$...**do**
  obtain direction V_k by semi-smooth newton solving the subproblem;
  set $\alpha = 1$
  while $F(Retr_{X_k}(\alpha V_k)) > F(X_k) - \frac{\alpha \| V_k \|_F^2}{2t}$ **do**
    $\alpha = \lambda\alpha$
    linesearchflag=1
  end while
  set $X_{k+1} = Retr_{X_k}(\alpha V_k)$
  if linesearchflag=1 **then**
  $t = \tau t$
  else $t = \max\{1/L, t/\tau\}t$
  end if
end for

[8]Chen, Shixiang, et al. "Proximal gradient method for nonsmooth optimization over the stiefel manifold[J]." Siam Journal on Optimization, vol. 30, no. 1, 2020, pp. 210 – 239.

Stage work summary

Algorithm 10 (Non-monotone line search with BB stepsize ManPG)

Input:initial point $X_0 \in M$,$\gamma \in (0,1)$,stepsize $t > 0, \eta, \delta \in 0,1$, $C_0 = F(X_0)$, $Q_0 = 1$.
for $k = 0, 1$...**do**
  obtain direction V_k by semi-smooth newton solving the subproblem;
  while $\| V_k \|_F > \epsilon$ **do**
    while $F(Retr_{X_k}(\alpha V_k)) > C_k + \frac{\alpha \| V_k \|_F^2}{2t}$ **do**
      $\alpha = \delta \alpha$;
    end while
    set $X_{k+1} = Retr_{X_k}(\alpha V_k)$;
    Calculate $Q_{k+1} = \eta Q_k + 1$;
    $C_{k+1} = \frac{\eta Q_k C_k + F(X_{k+1})}{Q_{k+1}}$;
    Choose $\alpha = |\alpha_k^{BB1}|$ or well $\alpha = |\alpha_k^{BB2}|$;
    Set $\alpha = \max(\min(\alpha, \alpha_{\max}), \alpha_{\min})$.
  end while
  $k = k + 1$;
end for

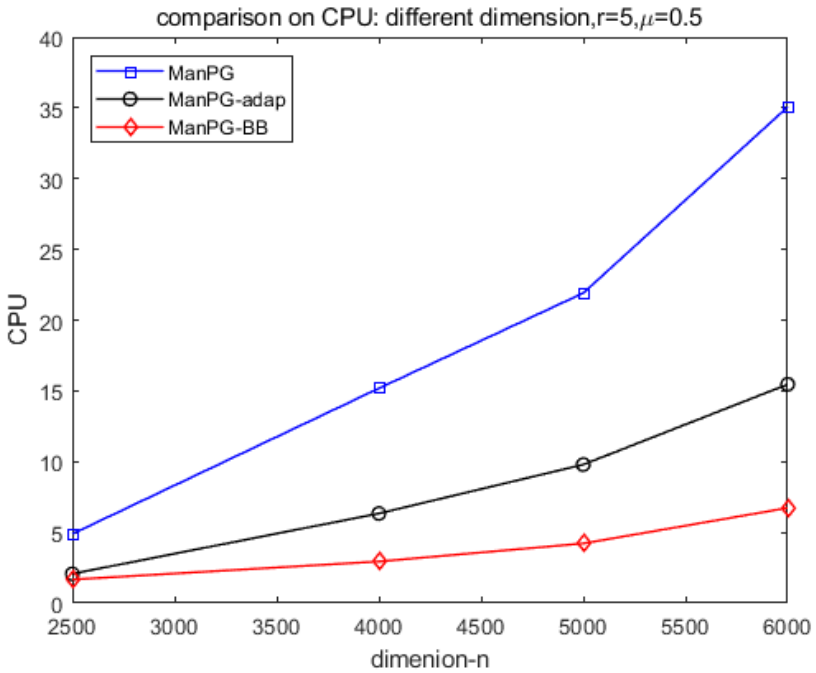
$\alpha_k^{BB1} = \frac{\| S_k \|_F^2}{tr(S_k^T R_k)}$ and $\alpha_k^{BB2} = \frac{tr(S_k^T R_k)}{\| R_k \|_F^2}$;where $S_k = X_{k+1} - X_k, R_k = V_{k+1} - V_k$

[9]Raydan, M.: The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM J. Optim. 7, 26 – 33 (1997).
[10]Zhang H , Hager W W . A nonmonotone line search technique and its application to unconstrained optimization[J]. SIAM Journal on Optimization, 2004, 14(4):1043-1056.

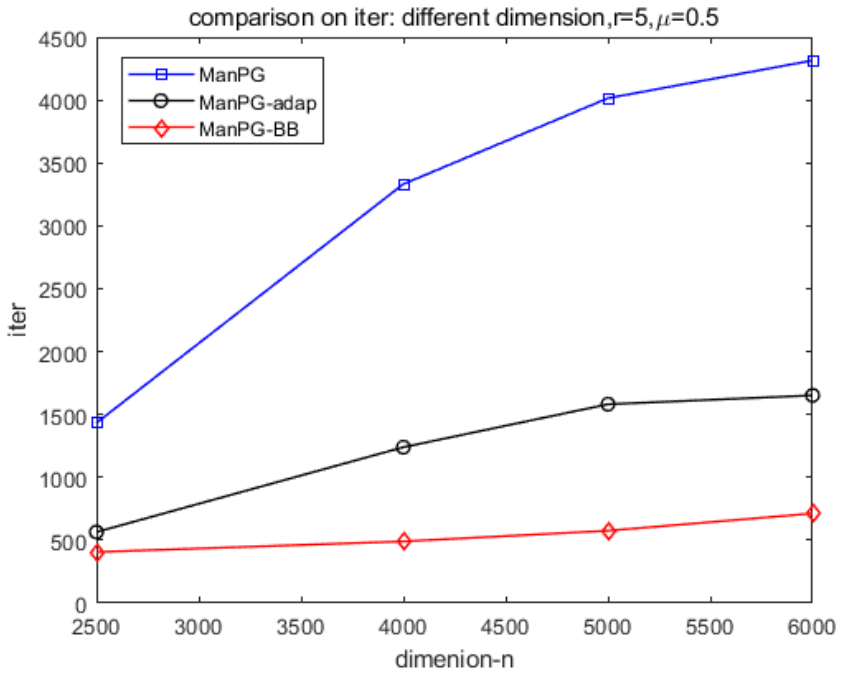
Numerical experiments

For Sparse PCA problem

$$\min_{X\in\mathbb{R}^{n\times p}} \; -\frac{1}{2}\,\mathrm{tr}\left(X^\top MX\right)+\mu\|X\|_1 \quad \text{s.t.} \quad X^\top X = I_p \tag{32}$$



(a) CPU

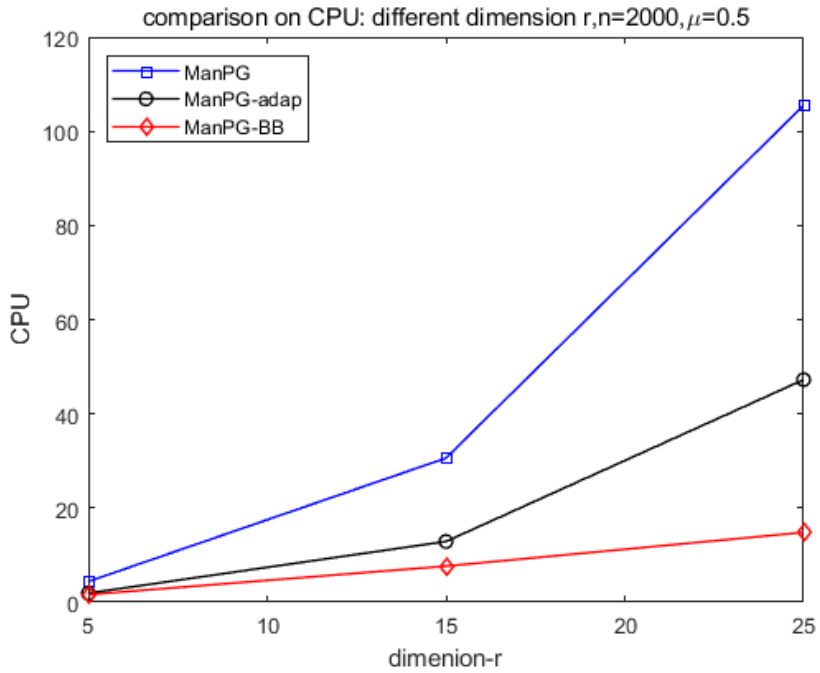


(b) Iteration

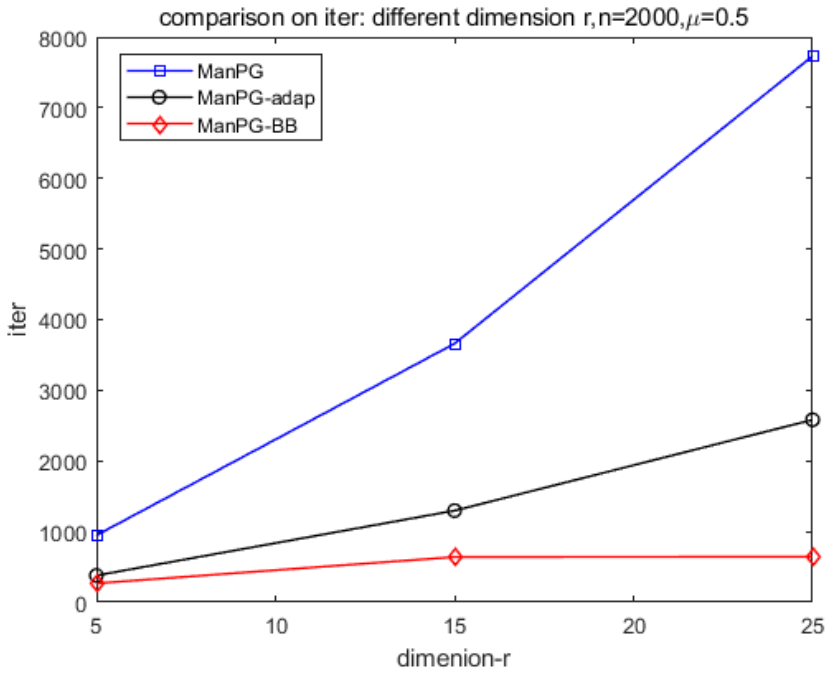
Figure 2: Comparison on SPCA problem with different n.

Numerical experiments

For Sparse PCA problem



(a) CPU



(b) Iteration

Figure 3: Comparison on SPCA problem with different p.

References


Chen S , Ma S , So M C , et al. Proximal gradient method for nonsmooth optimization over the stiefel manifold[J]. SIAM Journal on Optimization, 2020, 30(1):210-239.


N. Xiao, X. Liu and Y. Yuan, A class of smooth exact penalty function methods for optimization problems with orthogonality constraints, Optimization Methods and Software.2020 (DOI:10.1080/10556788.2020.1852236).


Nachuan Xiao, Xin Liu, Ya-xiang Yuan. Exact penalty function for $\ell_{2,1}$ norm minimization over the stiefel manifold.2020.


W. Chen, H. Ji, and Y. You, An augmented lagrangian method for ℓ_1 regularized optimization problems with orthogonality constraints[J], SIAM J. Sci. Comput., 38 (2016), pp. B570–B592.


Wang, Bokun, et al. “Riemannian stochastic proximal gradient methods for nonsmooth optimization over the stiefel manifold.” ArXiv Preprint ArXiv:2005.01209, 2020.

References


Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. Journal of Scientific Computing, 58(2):431 – 449, 2014.


Hu, Jiang, et al. "A brief introduction to manifold optimization." Journal of the Operations Research Society of China, vol. 8,no. 2, 2020, pp. 199 – 248.


Bin Gao, Xin Liu, Ya-xiang Yuan. Parallelizable algorithms for optimization problems with orthogonality constraints[J]. SIAM Journal on Scientific Computing, 41-3 (2019), A1949 – A1983.


Francisco, J.B., Gonalves, D.S., Bazán, F.S.V. et al. Nonmonotone inexact restoration approach for minimization with orthogonality constraints[J]. Numer Algor (2020).


Oviedo, Harry, et al. "Two adaptive scaled gradient projection methods for stiefel manifold constrained optimization[J]." Numerical Algorithms, 2020, pp. 1 – 21.

Thank you!