Structured Quasi-Newton Methods for Optimization with Orthogonality Constraints

Jiang Hu

Peking University

Joint work with Bo Jiang, Lin Lin, Zaiwen Wen and Yaxiang Yuan

The 12th International Conference on Numerical Optimization and Numerical Linear Algebra

Apr. 17, 2019

Outline

- Related applications
- Preliminaries on Riemannian optimization
- Adaptive regularized quasi-Newton method
- Numerical experiments

Outline

Electronic structure calculation: Kohn-Sham

- Kohn-Sham total energy minimization

$$\min_{X \in \mathbf{C}^{n\times p}} \quad E_{ks}(X) \quad \text{s.t.} \quad X^*X = I_p,$$

$$E_{ks}(X) := \frac{1}{4}\mathrm{tr}(X^*LX) + \frac{1}{2}\mathrm{tr}(X^*V_{ion}X) + \frac{1}{2}\sum_l\sum_i |x_i^* w_l|^2 + \frac{1}{4}\rho^\top L^\dagger \rho + \frac{1}{2}e^\top \epsilon_{xc}(\rho),$$

where $\rho := \rho(X) = \mathrm{diag}(XX^*)$.

- The Euclidean gradient and Hessian

$$\nabla E_{ks}(X) = H_{ks}(\rho)X, \quad \nabla^2 E_{ks}(X)[U] = H_{ks}(\rho)U + R(X)[U],$$

$$H_{ks}(\rho) := \frac{1}{2}L + V_{ion} + \sum_l w_l w_l^* + \mathrm{Diag}((\Re L^\dagger)\rho) + \mathrm{Diag}(\mu_{xc}(\rho)^* e),$$

$$R(X)[U] := \mathrm{Diag}\left((\Re L^\dagger + \frac{\partial^2 \epsilon_{xc}}{\partial \rho^2}e)(\bar{X}\odot U + X \odot \bar{U})e\right)X.$$

- First-order optimality: $H_{ks}(\rho)X = X\Lambda,\ X^*X = I_p$.

Fock exchange operator

- The operator $\mathcal{V}(\cdot) : \mathbf{C}^{n\times n} \to \mathbf{C}^{n\times n}$ is a fourth-order tensor
- For any $D_1, D_2 \in \mathbf{C}^{n\times n}$: $\langle \mathcal{V}(D_1), D_2\rangle = \langle \mathcal{V}(D_2), D_1\rangle$ and

$$\langle \mathcal{V}(D_1 + D_2), D_1 + D_2\rangle = \langle \mathcal{V}(D_1), D_1\rangle + 2\langle \mathcal{V}(D_1), D_2\rangle + \langle \mathcal{V}(D_2), D_2\rangle .$$

- Computing $\mathcal{V}(U)$ is very expensive since it needs to perform the multiplication between a $n \times n \times n \times n$ fourth-order tensor and a n-by-n matrix.
- The corresponding Fock energy is defined as

$$E_f(X) := \frac{1}{4}\langle \mathcal{V}(XX^*)X, X\rangle = \frac{1}{4}\langle \mathcal{V}(XX^*), XX^*\rangle . \tag{1}$$

Electronic structure calculation: Hartree-Fock

Hartree-Fock total energy minimization

$$\min_{X \in \mathbf{C}^{n\times p}} \quad E_{hf}(X) \quad \text{s.t.} \quad X^*X = I_p,$$

$$E_{ks}(X) := \frac{1}{4}\mathrm{tr}(X^*LX) + \frac{1}{2}\mathrm{tr}(X^*V_{ion}X) + \frac{1}{2}\sum_l \sum_i \zeta_l |x_i^* w_l|^2 + \frac{1}{4}\rho^\top L^\dagger \rho$$

$$E_{hf}(X) := E_{ks}(X) + E_f(X), \quad E_f(X) := \frac{1}{4}\langle V(D)X, X\rangle, \quad D := D(X) = XX^*$$

- The Euclidean gradient and Hessian

$$\nabla E_f(X) = V(D)X, \quad \nabla^2 E_f(X)[U] = V(D)U + V(XU^* + UX^*)X.$$

- First-order optimality

$$(H_{ks}(\rho) + V(D))X = X\Lambda, \; X^*X = I_p.$$

- $\nabla^2 E_f$ is much more expansive than $\nabla^2 E_{ks}$ due to the high computational cost of $V(D)X$.

Linear eigenvalue problem

- Linear eigenvalue problem

$$\min_{X \in \mathbb{R}^{n \times p}} f(X) := \frac{1}{2}\mathrm{tr}(X^\top (A + B)X) \quad \text{s.t.} \quad X^\top X = I_p,$$

- We assume that the multiplication of BX is much more expensive than that of AX.
- Euclidean gradient and Hessian

$$\nabla f(X) = (A + B)X, \quad \nabla^2 f(X)[U] = AU + BU.$$

Outline

Optimization with orthogonality constraints

Problem definition

$$\min_{X \in \mathbb{R}^{n \times p}} \quad f(X), \qquad \text{s.t.} \quad X^\top X = I_p,$$

where f is a differentiable function.

- The set of orthogonality matrices $\{X \in \mathbb{R}^{n \times p} \mid X^\top X = I_p\}$ is called the Stiefel manifold $\mathrm{St}(n, p)$.
- Applications: linear eigenvalue problem, electronic structure calculations, etc.

Riemannian optimization

Retraction

A retraction R on a manifold $\mathcal{M}$ at a point X is a mapping from the tangent space $T_X\mathcal{M}$ at X onto $\mathcal{M}$ satisfying

- $R_X(0_X) = X$, where 0_X denotes the zero tangent vector of $T_X\mathcal{M}$.
- $\frac{D}{dt}R_X(t\xi) = \xi$ for all $\xi \in T_X\mathcal{M}$.

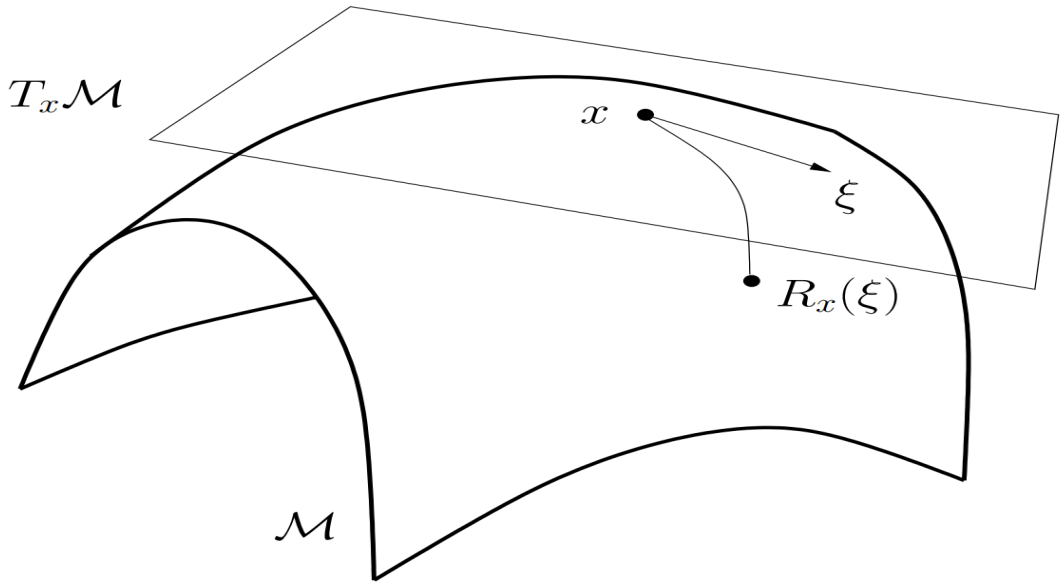


Figure: Absil et al. 2008

Riemannian optimization — Stiefel manifold

- Riemannian metric

$$\langle U, V\rangle := \mathrm{tr}(U^\top V),\ U, V \in T_X\mathcal{M}.$$

- The projection operator onto the tangent space is given by

$$\mathbf{P}_X(Z) = Z - X\mathrm{sym}(X^\top Z).$$

- Riemannian gradient and Hessian

$$\begin{aligned} \mathrm{grad} f(X) &= \mathbf{P}_X(\nabla f(X)), \\ \mathrm{Hess} f(X)[U] &= \mathbf{P}_X(\mathrm{Dgrad} f(X)[U]), \\ &= \mathbf{P}_X(\nabla^2 f(X)[U]) - U\mathrm{sym}(X^\top \nabla f(X)). \end{aligned}$$

- Riemannian optimization

$$X^{k+1} = R_{X^k}(\alpha_k \eta_k)$$

where α_k is the stepsize and $\eta_k \in T_{X^k}\mathcal{M}$ is a descent direction, e.g., $-\mathrm{grad} f(X^k)$ and $-\mathrm{Hess}^{-1} f(X^k)[\mathrm{grad} f(X^k)]$.

Outline

Existing Riemannian quasi-Newton method

- Focus on the whole approximation B^k to Riemannian Hessian

$$\mathrm{Hess} f(X^k) \ : \ T_{X^k}\mathcal{M} \to T_{X^k}\mathcal{M}.$$

- Riemannian BFGS method

$$B^{k+1} = \hat{B}^k - \frac{\hat{B}^k S^k ((\hat{B}^k)^* S^k)^\flat}{((\hat{B}^k)^* S^k)^\flat S^k} + \frac{Y^k (Y^k)^\flat}{(Y^k)^\flat S^k}, \ T_{X^{k+1}}\mathcal{M} \to T_{X^{k+1}}\mathcal{M}$$

where

$$\begin{aligned}
\hat{B}^k &= \mathbf{P}_{X^k}^{X^{k+1}} \circ B^k \circ (\mathbf{P}_{X^k}^{X^{k+1}})^{-1}, \text{change domain and range to } T_{X^{k+1}}\mathcal{M} \\
Y^k &= \beta_k^{-1} \mathrm{grad} f(X^{k+1}) - \mathbf{P}_{X^k}^{X^{k+1}} \mathrm{grad} f(X^k), \text{difference on } T_{X^{k+1}}\mathcal{M} \\
S^k &= \mathbf{P}_{X^k}^{X^{k+1}} \alpha_k \xi_k, \text{transport to } \mathrm{T}_{\mathrm{X}^{\mathrm{k+1}}}\mathcal{M}
\end{aligned}$$

with the last quasi-Newton direction $\xi_k \in T_{X^k}\mathcal{M}$ and stepsize α_k.

- $\mathbf{P}_{X^k}^{X^{k+1}} : T_{X^k}\mathcal{M} \to T_{X^{k+1}}\mathcal{M}$ is to transport the tangent vector from $T_{X^k}\mathcal{M}$ to $T_{X^{k+1}}\mathcal{M}$. β_k is a scalar (can be 1).

Existing Riemannian quasi-Newton method

- After getting B^{k+1}, the subproblem is

$$\xi_{k+1} := \arg \min_{\xi \in T_{X^{k+1}}\mathcal{M}} \quad \left\langle \operatorname{grad} f(X^{k+1}), \xi \right\rangle_{X^{k+1}} + \frac{1}{2} \left\langle B^{k+1}[\xi], \xi \right\rangle_{X^{k+1}}$$

- Do curvilinear search along xi_k to get stepsize α^k

$$X^{k+1} = R_{X^k}(\alpha_k \xi_k).$$

- To guarantee the fast local convergence
  - β_k and $\mathbf{P}_{X^k}^{X^{k+1}}$ should be chosen properly (satisfying locking condition[1]) to preserve curvature condition ($(Y^k)^\flat S^k > 0$).

[1] W. Huang, K. A. Gallivan, and P.-A. Absil, A Broyden class of quasi-Newton methods for Riemannian optimization, SIAM J. Optim., 25 (2015), pp. 16601685.

Adaptive regularized quasi-Newton method

- Riemannian Hessian of f:
  $\mathrm{Hess} f(X)[U] = \mathbf{P}_X(\nabla^2 f(X)[U]) - U\mathrm{sym}(X^\top \nabla f(X))$
- Keep the term $U\mathrm{sym}((X^k)^\top \nabla f(X^k))$ of lower computational cost, and construct an approximation B^k to expensive part $\nabla^2 f(X^k)$.
- After obtaining B^k, the subproblem is constructed as

$$\begin{cases} \min\ m_k(X) := \left\langle \nabla f(X^k), X - X^k \right\rangle + \frac{1}{2}\left\langle B^k[X - X^k], X - X^k \right\rangle + \frac{\sigma_k}{2}\|X - X^k\|^2, \\ \text{s.t. } X^T X = I_p. \end{cases}$$

- The Riemannian Hessian of $m_k(X)$ at X^k

$$\mathrm{Hess} m_k(X^k)[U] = \mathbf{P}_X(B^k[U]) - U\mathrm{sym}((X^k)^\top \nabla f(X^k)) + \sigma_k U.$$

- The vector transport is not needed since we are working the ambient Euclidean space.

Algorithm

- A modified conjugate gradient (CG) method
  Riemannian Newton equation:
  $$\mathrm{Hess} m_k(X^k)[\xi] = -\mathrm{grad} f(X^k) = \mathrm{grad}\, m_k(X^k)$$
  - Set $\xi_0 = 0, p_0 = -\mathrm{grad} f(X^k)$ and $i = 0$,
  - If negative curvature p_k is encountered, then
  $$\xi_k = \xi_{k-1} + \left\langle \mathrm{grad} f(X^k), p_j \right\rangle / \left\langle p_j, \mathrm{Hess} m_k(X^k)[p_j] \right\rangle,$$
  and return. Otherwise, do the normal truncated CG update ξ_k.
  - Do Armijo search along ξ^k to obtain a new trial point Z^k
- Choice of regularization parameter and updates
  - ratio: $\rho_k = \frac{f(Z^k) - f(X^k)}{m_k(Z^k)}$.
  - regularization parameter σ_k:

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k) & \text{if } \rho_k > \eta_2, \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \eta_1 \le \rho_k \le \eta_2, \\ (\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{otherwise.} \end{cases} \quad \begin{array}{l} \Rightarrow \boxed{X^{k+1} = Z^k} \\ \Rightarrow \boxed{X^{k+1} = Z^k} \\ \Rightarrow \boxed{X^{k+1} = X^k} \end{array}$$

where $0 < \eta_1 \le \eta_2 < 1$ and $1 < \gamma_1 \le \gamma_2$.

Construction of B^k with structured f

- Assume that $\nabla^2 f(X)$ takes a natural structure as

$$\nabla^2 f(X) = H^c(X) + H^e(X),$$

where the computational cost of $H^e(X)$ is much more expensive than that of $H^c(X)$.

- Quasi-Newton method with structure

$$B^k = H^c(X^k) + C^k,$$

where C^k is a quasi-Newton approximation to $H^e(X^k)$ with secant condition

$$C^k[S^k] = Y^k - H^c(X^k)[S^k],$$

where $S^k := X^k - X^{k-1}$ and $Y^k = \nabla f(X^k) - \nabla f(X^{k-1})$.

How to choose an initial quasi-Newton approximation?

- For a linear operator A of high computational cost, the limited-memory Nyström approximation[2] $\hat{A}$ is

$$\hat{A} := Y(Y^*\Omega)^\dagger Y^*,$$

  where $Y = A\Omega$ and Ω is a basis of a well-chosen subspace, e.g.,

$$\texttt{orth}(\{X^k, X^{k-1}, AX^k\}),\ \texttt{orth}(\{X^k, X^{k-1}, X^{k-2}, \ldots\}).$$

- The compressed operator $\hat{A}$ is of low rank, but consistent with A on the subspace spanned by Ω.
- Given some good approximation C_0^k of H^e, the Nyström approximation $\hat{C}_0^k$ can be utilized to further reduce the computational cost.
- More effective than the BB-type initialization (αI) in practice.

[2] Joel A Tropp, Alp Yurtsever, Madeleine Udell, and Volkan Cevher, Fixed-rank approximation of a positive-semidefinite matrix from streaming data, NIPS, 2017, pp. 1225-1234.

Application to electronic structure calculation

- Hartree-Fock total energy minimization

$$E_{hf}(X) := E_{ks}(X) + E_f(X), \quad E_f(X) := \frac{1}{4}\langle V(D)X, X\rangle .$$

- The structured update

$$H^c(X) = \nabla^2 E_{ks}(X), \quad H^e(X) = \nabla^2 E_f(X).$$

- Since $V(D^k)X^k$ can be obtained by computing the gradient $\nabla E_f(X^k)$, the Nyström approximation of V on $\{X^k\}$

$$\hat{V} := Y(Y^* X^k)^\dagger Y^*, \text{ with } Y = V(D^k)X^k.$$

- The subproblem

$$\begin{cases} \min m_k(X) := \left\langle \nabla f(X^k), X - X^k \right\rangle + \frac{1}{2}\left\langle B^k[X - X^k], X - X^k \right\rangle + \frac{\sigma_k}{2}\|X - X^k\|^2 \\ \text{s.t. } X \in \mathrm{St}(n,p). \end{cases}$$

Application to linear eigenvalue problem

- Linear eigenvalue problem

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) := \frac{1}{2}\mathrm{tr}(X^\top(A+B)X) \quad \text{s.t.} \quad X^\top X = I_p,$$

- We assume that the multiplication of BX is much more expensive than that of AX.
- The structured update

$$H^c(X) = A, \quad H^e(X) = B.$$

- Limited-memory Nyström approximation of B on subspace $W^k := \mathrm{span}\{X^{k-1}, X^k\}$

$$B^k := Y(Y^\top W^k)^\dagger Y^\top, \text{ with } Y = BW^k.$$

- The subproblem

$$m_k(X) := \frac{1}{2}\mathrm{tr}(X^\top(A+B^k)X) + \frac{\sigma_k}{4}\|XX^\top - X^k(X^k)^\top\|_F^2$$

Inexact conditions for convergence

Inexact conditions for solving subproblem:

(C1) Inexact condition for global convergence:

$$m_k(Z^k) \leq -c\|\mathrm{grad} f(X^k)\|_F^2$$

with some positive constant c.

(C2) Inexact condition for local convergence:

$$\|\mathrm{grad}\, m_k(Z^k)\|_F \leq \theta^k \|\mathrm{grad} f(X^k)\|_F$$

with parameter $\theta^k := \min\left\{0.1 * \|\mathrm{grad} f(X^k)\|, \|\mathrm{grad} f(X^k)\|_F^{1+\gamma}\right\} (\gamma > 0)$.

Global convergence

Assumptions:

(A1) The gradient ∇f is Lipschitz continuous on the convex hull of the manifold $\mathrm{conv}(\mathrm{St}(n,p))$, i.e., there exists $L_f > 0$ such that

$$\|\nabla f(X) - \nabla f(Y)\| \le L_f \|X - Y\|, \quad \forall\, X, Y \in \mathrm{conv}(\mathrm{St}(n,p)).$$

(A2) There exists $\kappa_H > 0$ such that $\|B^k\| \le \kappa_H$ for all $k = 1, 2, \ldots$.

Theorem

Suppose that the assumptions (A1)–(A2) and condition (C1) hold. Then, either

$$\mathrm{grad} f(X^\ell) = 0 \ \text{ for some } \ \ell \ge 0 \quad \text{or} \quad \liminf_{k\to\infty} \|\mathrm{grad} f(X^k)\| = 0.$$

Local convergence

Assumptions:

(B1) The sequence $\{X^k\}$ converges to X_* with $\mathrm{grad} f(X_*) = 0$.

(B2) The Euclidean Hessian $\nabla^2 f$ is continuous on $\mathrm{conv}(\mathcal{M})$.

(B3) The Riemannian Hessian $\mathrm{Hess} f(X)$ is positive definite at X_*.

(B4) The Hessian approximation B^k satisfies

$$\frac{\|(B^k - \nabla^2 f(X^k))[Z^k - X^k]\|_F}{\|Z^k - X^k\|_F} \to 0, \ k \to \infty.$$

Theorem

Suppose that the assumptions (B1)-(B4) and conditions (C1)-(C2) hold. Then the sequence $\{X^k\}$ converges q-superlinearly to X_*.

Outline

Numerical results: Hartree-Fock energy minimization

Hartree-Fock total energy minimization

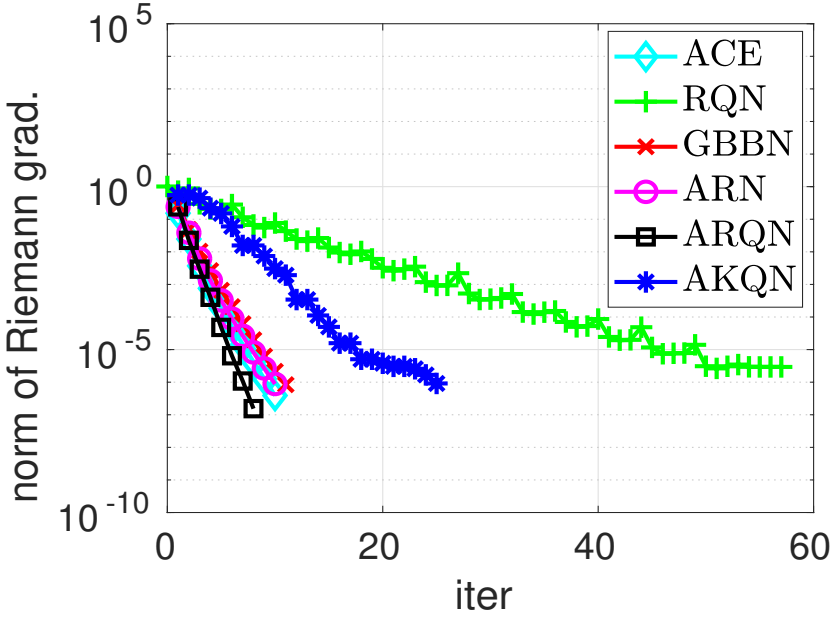
$$\min_{X \in \mathbf{C}^{n\times p}} \quad E_{hf}(X) := E_{ks}(X) + E_f(X) \quad \text{s.t.} \quad X^*X = I_p.$$

- ACE: existing two-level nested self-consistent field iteration with the Nyström approximation.
- ARQN(our method): keeps the Hessian $\nabla^2 E_{ks}$ and construct an approximation to $\nabla^2 E_f$.
- AKQN(our method): only keeps the linear operator part in $\nabla^2 E_{ks}$.
- ARN(our method): the only difference to ARQN is to set $B^k = \nabla^2 E_{ks}(X^k) + \hat{V}(D^k)$ directly, where $\hat{V}(D^k)$ is a Nyström approximation to $\nabla^2 E_f(X^k)$.
- GBBN: the only difference to ARN is that the subproblem is solved by GBB.
- RQN: Riemannian limited-memory BFGS method from Manopt

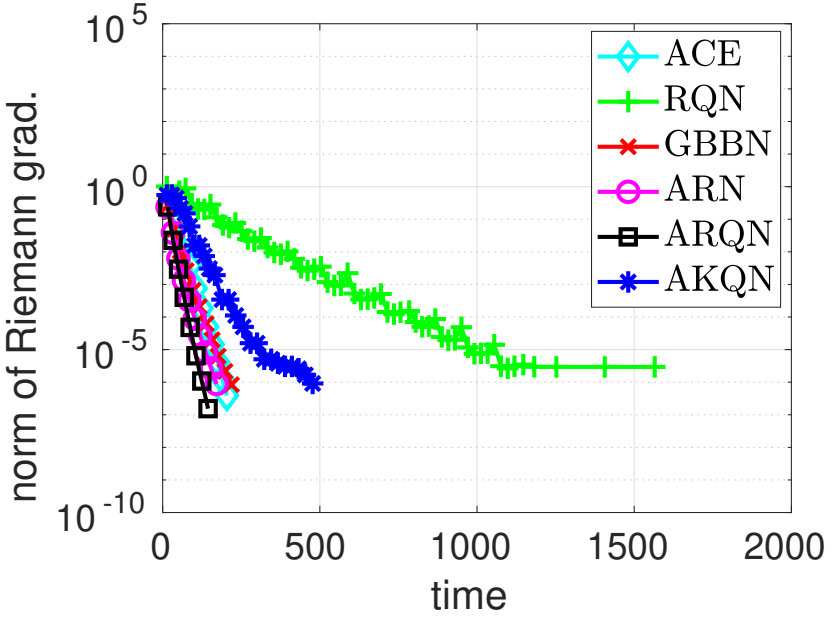
Numerical results: Hartree-Fock energy minimization

Table: Numerical results on HF total energy minimization.

Solver	fval	nrmG	its	time	fval	nrmG	its	time
	glutamine				graphene30			
ACE	-1.04525e+2	3.9e-7	10(3.0)	229.6	-1.87603e+2	8.6e-7	58(4.2)	15182.3
GBBN	-1.04525e+2	8.4e-7	11(13.3)	256.9	-1.87603e+2	8.6e-7	32(76.0)	22678.9
ARN	-1.04525e+2	8.8e-7	10(9.5)	209.5	-1.87603e+2	9.0e-7	45(35.6)	14941.2
ARQN	-1.04525e+2	1.5e-7	8(10.1)	182.9	-1.87603e+2	7.6e-7	15(26.5)	5873.2
AKQN	-1.04525e+2	9.1e-7	25(6.0)	515.7	-1.87603e+2	9.5e-7	62(7.5)	18986.5
RQN	-1.04525e+2	2.9e-6	57	1532.8	-1.87603e+2	1.5e-5	110	39057.2
	gaas				si40			
ACE	-2.93496e+2	8.8e-7	29(2.9)	343.8	-1.65698e+2	9.2e-7	29(4.5)	30256.4
GBBN	-2.93496e+2	9.3e-7	34(35.3)	659.3	-1.65698e+2	8.6e-7	24(43.9)	34846.2
ARN	-2.93496e+2	9.6e-7	31(20.4)	468.7	-1.65698e+2	8.0e-7	22(22.1)	21181.3
ARQN	-2.93496e+2	3.3e-7	10(28.0)	199.5	-1.65698e+2	2.8e-7	12(37.8)	15369.5
AKQN	-2.93496e+2	4.6e-7	22(18.4)	347.1	-1.65698e+2	9.2e-7	87(7.9)	89358.8
RQN	-2.93496e+2	1.0e-6	126	2154.1	-1.65698e+2	6.1e-6	156	181976.8



(a) norm of Riemannian gradients



(b) norm of Riemannian gradients

Figure: Comparisons of different algorithms on "glutamine" of HF total energy minimization.

Numerical results: linear eigenvalue problem

Linear eigenvalue problem

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) := \frac{1}{2}\mathrm{tr}(X^\top(A+B)X) \quad \text{s.t.} \quad X^\top X = I_p, \tag{2}$$

- We assume that the multiplication of BX is much more expensive than that of AX.

- EIGS: built-in function in MATLAB.
- LOBPCG: locally optimal block preconditioned conjugate gradient method.
- ASQN (our method): keeps A but uses the limited-memory Nyström approximation of B on subspace $\mathrm{span}\{X^{k-1}, X^k\}$.
- ACE: keeps A but uses the limited-memory Nyström approximation of B on subspace $\mathrm{span}\{X^k\}$.

Numerical results: linear eigenvalue problem

Settings:

- $A = \text{randn}(n, n)$; $A = (A + A^\top)/2$;
- $B = 0.01\text{randn}(n, n)$; $B = (B + B^\top)/2$; $B = B - \lambda_{\max}(B)$;
- we compute the multiplication BX using $\frac{1}{19}\sum_{i=1}^{19} BX$

Table: Numerical results on random matrices

	AV/BV	err	time	AV/BV	err	time
n	8000			10000		
EIGS	538/538	8.7e-11	131.9	981/981	8.8e-11	327.3
LOBPCG	1996/1996	9.9e-11	336.7	2440/2440	9.7e-11	763.8
ASQN	2706/150	8.9e-11	29.8	2920/150	9.7e-11	50.2
ACE	4537/450	9.8e-11	66.3	4554/400	9.6e-11	99.4
$n = 5000$						
p	30			50		
EIGS	660/660	3.0e-11	62.8	879/879	1.6e-12	83.6
LOBPCG	4458/4458	1.0e-10	217.6	5766/5766	9.5e-11	186.7
ASQN	5315/420	9.8e-11	11.4	7879/650	9.8e-11	17.8
ACE	9701/1530	9.4e-11	23.0	21664/4450	1.0e-10	50.9

Numerical results: linear eigenvalue problem

Settings:

- $A = \text{gallery}(\text{'wathen'}, 5s, 5s)$
- $B = 0.01\text{randn}(n, n);\ B = (B + B^\top)/2;\ B = B - \lambda_{\max}(B);$

Table: Numerical results on sparse matrices

	AV/BV	err	time	AV/BV	err	time
$p = 10$						
s	11			12		
EIGS	1882/1882	1.5e-07	58.9	1463/1463	9.6e-11	65.4
LOBPCG	4282/4282	9.5e-11	136.0	4089/4089	9.9e-11	190.6
ASQN	8327/240	9.6e-11	16.7	6910/220	9.3e-11	17.5
ACE	15323/1060	9.7e-11	38.9	17907/2010	1.7e-08	65.5
$s = 12$						
p	50			60		
EIGS	1743/1743	7.3e-11	69.1	2122/2122	1.6e-11	86.7
LOBPCG	12288/12288	1.4e-09	168.4	15716/15716	1.1e-08	199.5
ASQN	21330/1300	9.3e-11	53.6	26343/1620	9.7e-11	71.8
ACE	49165/10050	2.9e-06	110.1	62668/12060	2.3e-08	134.0

Summary

- We propose a quasi-Newton method without vector transport.
- The structured update of the quasi-Newton approximation is presented by utilizing the structure of the objective f.
- The limited-memory Nyström approximation is investigated to obtain some good initial approximation for the exact Hessian.
- Numerical experiments compared with the state-of-art methods show the effectiveness of our method.
- More information can be found in `https://arxiv.org/abs/1809.00452`

Many Thanks For Your Attention!