Structured Quasi-Newton Method for Optimization with Orthogonality Constraints

Jiang Hu jianghu@pku.edu.cn

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

Beijing International Center for Mathematical Research, Peking University

Introduction

- Optimization with unitary matrices

$$\min_{X\in\mathbb{C}^{n\times p}} \quad f(X) \quad \text{s.t. } X^*X = I_p, \tag{1}$$

where $f(X):\mathbb{C}^{n\times p}\to\mathbb{R}$ is a $\mathbb{R}$-differentiable function.

- Assume that the Euclidean Hessian $\nabla^2 f(X)$ takes a structure

$$\nabla^2 f(X) = \mathcal{H}^{\mathrm{c}}(X) + \mathcal{H}^{\mathrm{e}}(X), \tag{2}$$

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

- When f is a summation of functions whose full Hessian are expensive to be evaluated or even not accessible.
- Hartree-Fock total energy minimization in electronic structure calculation.

Applications

Linear eigenvalue problem

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

when the multiplication of BX is much more expensive than that of AX.

- The streaming model — a series of linear updates $A \leftarrow \theta_1 A + \theta_2 H$
- Linear subproblem of SCF for electronic structure calculation

Hartree-Fock total energy minimization

- The discretized Kohn-Sham density functional

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

- $X = [x_1,\ldots,x_p]\in\mathbb{C}^{n\times p}$, with $X^*X = I_p$
- $\rho := \rho(X) = \mathrm{diag}(XX^*)$
- A hybrid exchange-correlation operator to account for the electron-electron interaction $\mathcal{V}(\cdot):\mathbb{C}^{n\times n}\to\mathbb{C}^{n\times n}$, which is usually a fourth-oder tensor and hence is of large computational cost.
- Hartree-Fock total energy minimization

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

where $E_{\mathrm{f}}(X) := \frac{1}{4}\langle\mathcal{V}(XX^*)X, X\rangle = \frac{1}{4}\langle\mathcal{V}(XX^*), XX^*\rangle$

Contributions

Regarding the constraints as the Stiefel manifold, the Riemannian Hessian $\mathrm{Hess} f(X)$ is with following structure:

$$\mathrm{Hess} f(X)[\xi] = \mathrm{Proj}_X(\nabla^2 f(X)[\xi] - \xi\mathrm{sym}(X^*\nabla f(X))), \tag{3}$$

where $\xi\in T_X := \{\xi\in\mathbb{C}^{n\times p} : X^*\xi + \xi^*X = 0\}$, projection $\mathrm{Proj}_X(Z) := Z - X\mathrm{sym}(X^*Z)$ and $\mathrm{sym}(A) := (A+A^*)/2$.

- From the structure (3), we approximate Euclidean Hessian $\nabla^2 f(X)$ instead of the full Riemannian Hessian $\mathrm{Hess} f(X)$ directly, but keep the remaining parts $\xi\mathrm{sym}(X^*\nabla f(X))$ and $\mathrm{Proj}_X(\cdot)$.
- By further taking advantage of the structure (2) of f, we develop a quasi-Newton approach to construct an approximation to the expensive part $\mathcal{H}^{\mathrm{e}}$ while preserving the cheap part $\mathcal{H}^{\mathrm{c}}$. This kind of structured approximation usually yields a better property than the approximation constructed by the vanilla quasi-Newton method.
- For the construction of an initial approximation of $\mathcal{H}^{\mathrm{e}}$, we also investigate a limited-memory Nyström approximation, which gives a subspace approximation of a known good but still complicated approximation of $\mathcal{H}^{\mathrm{e}}$.
- When the subproblems are solved to certain accuracy, both global and local q-superlinear convergence can be established under certain mild conditions.
- The proposed algorithms perform comparably well with the state-of-art methods in linear eigenvalue problem and electronic structure calculation.

A Structured Quasi-Newton Approach

Structured quasi-Newton approximation

- Construct an approximation $\mathcal{E}^k$ to $\mathcal{H}^{\mathrm{e}}(X^k)$
- Keep the cheaper part $\mathcal{H}^{\mathrm{c}}(X^k)$, an approximation to $\nabla^2 f(X^k)$

$$\mathcal{B}^k = \mathcal{H}^{\mathrm{c}}(X^k) + \mathcal{E}^k$$

- $\mathcal{E}^k$ is an approximation to $\mathcal{H}^{\mathrm{e}}(X^k)$
- To ensure $\mathcal{B}^k[S^k] = Y^k$, the secant condition for $\mathcal{E}^k$

$$\mathcal{E}^k[S^k] = Y^k - \mathcal{H}^{\mathrm{c}}(X^k)[S^k],$$

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

- Utilize limited-memory symmetric rank-one update to construct $\mathcal{E}^k$ satisfying the following secant equation

Limited-memory Nyström approximation

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

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

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

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

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

A structured quasi-Newton method

- Objective for subproblem (Approximate f in Euclidean space)

$$m_k(X) := \left\langle g_k, X - X^k\right\rangle + \frac{1}{2}\left\langle\mathcal{B}^k[X-X^k], X-X^k\right\rangle + \frac{\tau_k}{2}d(X,X^k)$$

with $g_k := \nabla f(X^k)$.

- Keep the constraints and construct the subproblem as

$$\min_{X\in\mathbb{C}^{n\times p}} \quad m_k(X) \quad \text{s.t.} \quad X^*X = I. \tag{4}$$

- τ_k is the regularization parameter and $d(X,X^k)$ is a proximal term to guarantee the convergence.
- Set $d(X,X^k) = \|X-X^k\|_{\mathsf{F}}^2$, the Riemannian Hessian of $m_k(X)$

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

- The vector transport is not needed since we are working the ambient Euclidean space.
- A modified conjugate gradient (CG) method

  Newton's equation

$$\mathrm{Hess} m_k(X^k)[\xi_k] = -\mathrm{grad} f(X^k).$$

  – 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 to obtain the direction ξ_k.

  – Do Armijo search along ξ^k to obtain a new trial point Z^k

- Choice of regularization parameter and updates

  – Ratio:

$$\rho_k = \frac{f(Z^k) - f(X^k)}{m_k(Z^k)}. \tag{5}$$

  – Regularization parameter τ_k:

$$\tau_{k+1}\in\begin{cases}(0,\tau_k] & \text{if } \rho_k > \eta_2, \quad \Rightarrow \boxed{X^{k+1} = Z^k}\\ [\tau_k,\gamma_1\tau_k] & \text{if } \eta_1\le\rho_k\le\eta_2, \quad \Rightarrow \boxed{X^{k+1} = Z^k}\\ [\gamma_1\tau_k,\gamma_2\tau_k] & \text{otherwise.} \quad \Rightarrow \boxed{X^{k+1} = X^k}\end{cases} \tag{6}$$

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

Algorithm 1: A structured quasi-Newton method

Input: initial guess $X^0\in\mathbb{C}^{n\times p}$ with $(X^0)^*X^0 = I_p$ and $\tau_0 > 0$, choose $0<\eta_1\le\eta_2<1$, $1<\gamma_1\le\gamma_2$, set $k=0$.

while *stopping conditions not met* **do**

- Consturct subproblem and Use modified CG method to compute a new trial point Z^k.
- Compute the ratio ρ_k via (5).
- Update X^{k+1} from the trial point Z^k based on (6).
- Update τ_k according to (6).
- $k\leftarrow k+1$.

Convergence

Assumption 1. *Let $\{X^k\}$ be generated by Algorithm 1. We assume:*

(A.1) The gradient ∇f is Lipschitz continuous on the convex hull of the Stiefel manifold $\mathcal{M} := \{X\in\mathbb{C}^{n\times p} \mid X^*X = I_p\}$ – denoted by $\mathrm{conv}(\mathcal{M})$, i.e., there exists $L_f > 0$ such that

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

(A.2) There exists $\kappa_H > 0$ such that $\|\mathcal{B}_k\|\le\kappa_H$ for all $k\in\mathcal{N}$.

$$\|\nabla^2 f(X_k)\|\le\kappa_F, \forall\, k\in\mathcal{N}.$$

The inexact conditions for the subproblem (4) (with quadratic or cubic regularization) can be chosen as

$$m_k(Z^k) \le -\frac{a}{b+\tau_k}\|\mathrm{grad} f(X^k)\|_{\mathsf{F}}^2 \tag{7}$$

$$\|\mathrm{grad} m_k(Z^k)\|_{\mathsf{F}} \le \theta^k\|\mathrm{grad} f(X^k)\|_{\mathsf{F}} \tag{8}$$

where a,b,c are positive constants and $\theta^k := \min\{1, \|\mathrm{grad} f(X^k)\|_{\mathsf{F}}^c\}$

Theorem 2. *Suppose that the Assumptions (A.1)-(A.2) hold and let $\{f(X^k)\}$ be bounded from below. Then, either*

$$\mathrm{grad} f(X^\ell) = 0 \text{ for some } \ell > 0 \quad \text{or} \quad \lim_{k\to\infty}\|\mathrm{grad} f(X^k)\|_{\mathsf{F}} = 0.$$

Assumption 3. *Let $\{X^k\}$ be the sequence generated by Algorithm 1. We assume*

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

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

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

(B4) The Hessian approximation $\mathcal{B}^k$ satisfies

$$\frac{\|(\mathcal{B}^k - \nabla^2 f(X^k))[Z^k - X^k]\|_{\mathsf{F}}}{\|Z^k - X^k\|_{\mathsf{F}}} \to 0,\ k\to\infty. \tag{9}$$

Theorem 4. *Suppose that the conditions (B1)-(B4) and (8) hold. Then the sequence $\{X^k\}$ converges q-superlinearly to X_*.*

Linear Eigenvalue problem

$$\min_{X\in\mathbb{R}^{n\times p}} f(X) := \frac{1}{2}\mathrm{tr}(X^\top(A+B)X) \quad \text{s.t.} \quad X^\top X = I_p, \tag{10}$$

where $A, B\in\mathbb{R}^{n\times n}$ are symmetric matrices.

- The limited-memory Nyström approximation $\hat{B}^k$ on $\mathrm{orth}[X^{k-1}, X^k]$
- New $m_k(X)$ for subproblem

$$m_k(X) := \frac{1}{2}\mathrm{tr}(X^\top(A+\hat{B}^k)X) + \frac{\tau_k}{4}\|XX^\top - X^k(X^k)^\top\|_{\mathsf{F}}^2$$

- Numerical tests

$$A = \mathrm{randn}(n,n);\ A = (A+A^\top)/2;$$

$$B = 0.01\mathrm{rand}(n,n);\ B = (B+B^\top)/2;\ B = B - \lambda_{\min}(B)I_n;\ B = -B,$$

In our implementation, we compute the multiplication BX using $\frac{1}{m}\sum_{i=1}^m BX$, where m is chosen to be 19.

	#Av/#A/#Bv/#B	err	time	B-time	#Av/#A/#Bv/#B	err	time	B-time
	$p=10$							
n	8000				10000			
EIGS	538/529/538/529	8.7e-11	70.6	66.6	981/972/981/972	8.8e-11	153.8	144.8
LOBPCG	1996/314/1996/314	9.9e-11	134.0	57.2	2440/387/2440/387	9.7e-11	287.4	122.5
ASQN	2706/567/150/15	8.9e-11	11.2	2.8	2920/581/150/15	9.7e-11	17.8	5.4
ACE	4537/1162/450/45	9.8e-11	26.1	9.8	4554/951/400/40	9.6e-11	35.3	14.1
	$n=5000$							
p	30				50			
EIGS	660/631/660/631	3.0e-11	47.4	45.2	879/830/879/830	1.6e-12	47.7	44.6
LOBPCG	4412/707/4412/707	9.7e-11	111.2	56.1	5766/542/5766/542	9.5e-11	97.0	40.0
ASQN	5315/636/420/14	9.8e-11	7.9	1.3	7879/711/650/13	9.8e-11	12.6	1.8
ACE	9701/1173/1530/51	9.4e-11	15.8	4.6	21832/2270/4500/90	9.7e-11	41.4	13.2

"#Av" and "#Bv" denote the total number of matrix-vector multiplications (MV), counting each operation $AV, BV\in\mathbb{R}^{n\times p}$ as p MVs. The labels "#A" and "#B" are the total number of calls of A and B. ACE also utilizes the Nyström approximation on X^k but has no convergence guarantee for general B (only for semidefinite matrix B). ASQN is our method, which reduces the evaluations of BX instead of AX. Therefore, the convergence is accelerated.

Hartree-Fock Total Energy Minimization

- The gradient and Hessian of $E_{\mathrm{f}}(X)$

$$\nabla E_{\mathrm{f}}(X) = \mathcal{V}(XX^*)X,$$
$$\nabla^2 E_{\mathrm{f}}(X)[U] = \mathcal{V}(XX^*)U + \mathcal{V}(XU^* + UX^*)X.$$

- Since $\mathcal{V}(XX^*)X$ can be obtained from the gradient, we use its limited-Nyström approximation to serve as a initialization of the Quasi-Newton approximation.

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
ASQN	-1.04525e+2	1.5e-7	8(10.1)	182.9	-1.87603e+2	7.6e-7	15(26.5)	5873.2
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
ASQN	-2.93496e+2	3.3e-7	10(28.0)	199.5	-1.65698e+2	2.8e-7	12(37.8)	15369.5
RQN	-2.93496e+2	1.0e-6	126	2154.1	-1.65698e+2	6.1e-6	156	181976.8

ACE is an efficient method which utilizes the Nyström approximation but has no convergence guarantee. RQN is the Riemannian BFGS method from Manopt. ASQN is our method, which performs best in terms of both accuracy and time.