A Brief Introduction to Manifold Optimization

Zaiwen Wen

Beijing International Center For Mathematical Research
Peking University

http://bicmr.pku.edu.cn/~wenzw
wenzw@pku.edu.cn

Reference: J. Hu, X. Liu, Z. Wen, Y. Yuan, A Brief Introduction to Manifold Optimization, Journal of Operations Research Society of China

Model problem

$$\min_X \quad f(X), \quad X \in \mathcal{M}$$

Examples: Stiefel manifold, oblique manifold, Rank-p manifold, ...

- **important applications** from machine learning, material science and etc: eigenvalue decomposition, Quantum physics/chemisty, density functional theory, Bose-Einstein condensates, low rank nearest correlation matrix, Cryo-EM, phase retrieval, assignment matrix
- **Difficulty**: nonconvexty, multiple local minimizers/saddle points
- **Recent progress**
  - General first-order and second-order general algorithms/analysis
  - Algorithms/analysis for Linear and Nonlinear Eigenvalue Problem
  - Batch normalization from deep learning
  - Analysis of global optimal solution in maxcut type problems

Outline

1. Applications
2. Algorithms
3. Theory

Minimizing p-Harmonic Flows into Sphere

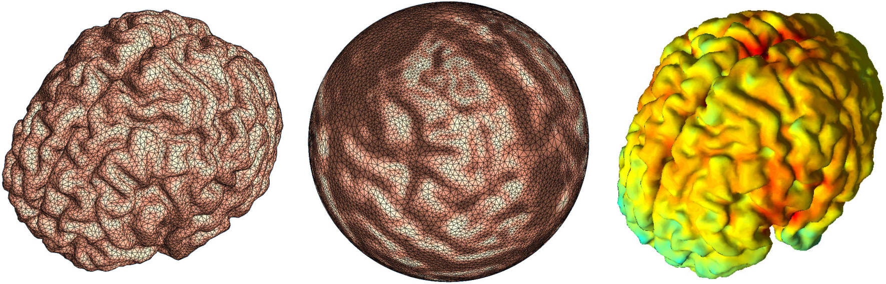


Figure: input surface; the conformal map; the surfaces are color coded by the corresponding u in the conformal factors.

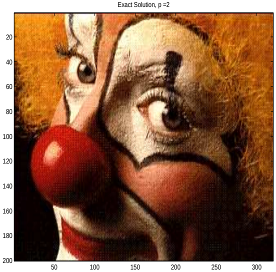
$$\min_{F=(f_1,f_2,f_3)} \quad \mathbf{E}(F) = \frac{1}{2}\int_{\mathcal{M}} \|\nabla_{\mathcal{M}} f_1\|^2 + \|\nabla_{\mathcal{M}} f_2\|^2 + \|\nabla_{\mathcal{M}} f_3\|^2 \mathrm{d}\mathcal{M}$$

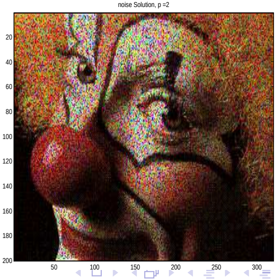
$$s.t. \quad \|F\| = \sqrt{f_1^2 + f_2^2 + f_3^2} = 1, \qquad \forall x \in \mathcal{M}$$

Minimizing p-Harmonic Flows into Sphere

$$\begin{cases} \min & \widehat{E}_p(\mathbf{U}) = \int_\Omega |\mathcal{D}\mathbf{U}(\mathbf{x})|_F^p \, \mathrm{d}\mathbf{x}, \\ s.t. & \mathbf{U} \in \{\mathbf{U} \in W^{1,p}(\Omega, \mathbb{R}^N) \,|\, |\mathbf{U}(\mathbf{x})| = 1 \; a.e.; \mathbf{U}|_{\partial\Omega} = \mathbf{n}_0\} \end{cases}$$

- Applications
  - directional diffusion, color image denoising, conformal mapping;
  - micromagnetics, i.e., describing magnetization patterns in ferromagnetic media (Minimizing the Landau-Lifshitz energy);
  - Computing liquid crystal's stable configuration





Maxcut type problems

- Original: binary variable $x_i \in \{-1, 1\}$.

$$\max_x \; \frac{1}{2}\sum_{i<j} w_{ij}(1 - x_i x_j), \quad \text{s.t. } x_i = \{\pm 1\},\; i = 1, \ldots, n.$$

- SDP relaxation: $xx^\top \to X \succeq 0$, drop $\mathrm{rank}(X) = 1$.

$$\max_X \quad \mathrm{tr}(CX), \quad \text{s.t.} \quad X_{ii} = 1,\; i = 1, \cdots, n,\; X \succeq 0.$$

- NLP: write $X = V^\top V$ where $V = [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n] \in \mathbb{R}^{p\times n}$

$$\max_{V \in \mathbb{R}^{p\times n}} \; \sum_{i,j} c_{ij} \boldsymbol{v}_i^\top \boldsymbol{v}_j, \quad \text{s.t. } \|\boldsymbol{v}_i\| = 1,\; i = 1, \ldots, n.$$

- Low-rank nearest correlation matrix estimation

$$\min \frac{1}{2}\left\| W \odot (V^\top V - C)\right\|_F^2, \quad \text{s.t. } \|\boldsymbol{v}_i\| = 1,\; i = 1, \ldots, n.$$

Partition Matrix from Community Detection

- For any partition $\cup_{a=1}^{k} C_a = [n]$, define the partition matrix X

$$X_{ij} = \begin{cases} 1, & \text{if } i,j \in C_a, \text{ for some } a, \\ 0, & \text{else }. \end{cases}$$

- Low rank solution

$$X = \begin{bmatrix} 1 & & & & & \\ & 1 & 1 & 1 & & \\ & 1 & 1 & 1 & & \\ & 1 & 1 & 1 & & \\ & & & & 1 & 1 \\ & & & & 1 & 1 \end{bmatrix} = \begin{bmatrix} 1 & & \\ & 1 & \\ & 1 & \\ & 1 & \\ & & 1 \\ & & 1 \end{bmatrix} \times \begin{bmatrix} 1 & & & & & \\ & 1 & 1 & 1 & & \\ & & & & 1 & 1 \end{bmatrix}$$

Modularity Maximization

- The modularity (*MEJ Newman, M Girvan, 2004*) is defined by

$$Q = \langle A - \frac{1}{2\lambda} dd^T, X \rangle$$

  where $\lambda = |E|$.

- The Integral modularity maximization problem:

$$\begin{aligned} \max \quad & \langle A - \tfrac{1}{2\lambda} dd^T, X \rangle \\ \text{s.t.} \quad & X \in \{0,1\}^{n\times n} \text{ is a partiton matrix.} \end{aligned}$$

  where $\lambda = |E|$.

- SDP Relaxation Yudong Chen, Xiaodong Li, Jiaming Xu

$$\begin{aligned} \max \quad & \langle A - \tfrac{1}{2\lambda} dd^T, X \rangle \\ \text{s.t.} \quad & X \succeq 0 \\ & 0 \le X_{ij} \le 1 \\ & X_{ii} = 1 \end{aligned}$$

Assignment matrix

- Optimization over permutation matrices (OptPerm)

$$\min_{X} f(X), \text{ s.t. }, X \in \Pi_n = \{X^\top X = I, \; X \geq 0\}. \tag{1}$$

- Quadratic assignment problem (QAP)

$$\min_{X \in \Pi_n} f(X) := \mathrm{tr}(A^\top X B X^\top), \tag{2}$$

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

- Graph matching problem

$$\min_{X \in \Pi_n} f(X) = \|AX - XB\|_F^2, \tag{3}$$

- $f(X) = \|AX - XB\|_F^2 = -\mathrm{tr}(A^\top X B X^\top) + \text{const}.$

Linear eigenvalue problem

Given a symmetric $n \times n$ real matrix A

- $k-$truncated decomposition ($k \ll n$):

$$AQ_k = Q_k\Lambda_k.$$

  - $\Lambda_k \in \mathbb{R}^{k\times k}$ contains k smallest/largest eigenvalues.
  - $Q_k \in \mathbb{R}^{n\times k}$ consists of the first/last k columns of Q.

- Trace minimization:

$$\min(\max) \quad \mathrm{tr}(X^\top AX), \text{ s.t. } X^\top X = I$$

- A fundamental tool for many emerging optimization
  - semidefinite program, Low-rank matrix completion, Robust principal component analysis, Sparse principal component analysis, Sparse inverse covariance matrix estimation, DFT, High dimensional data reduction

Electronic Structure Calculation

- Total energy minimization problem:

$$\min_{X^*X=I} E_{kinetic}(X) + E_{ion}(X) + E_{Hartree}(X) + E_{xc}(X) + E_{fock}(X),$$

where

$$\begin{aligned}
E_{kinetic}(X) &= \frac{1}{2}\mathrm{tr}(X^* L X) \\
E_{ion}(X) &= \mathrm{tr}(X^* V_{ion} X) \\
E_{Hartree}(X) &= \frac{1}{2}\rho(X)^\top L^\dagger \rho(X) \\
E_{xc}(X) &= \rho(X)^\top \mu_{xc}(\rho(X)) \\
\rho(X) &= \mathrm{diag}(D(X)), \quad D(X) = XX^* \\
E_{fock}(X) &= \langle V(D)X, X\rangle, \text{ fourth order tensor}
\end{aligned}$$

- Nonlinear eigenvalue problem (looks like the KKT condtions):

$$\begin{aligned}
H(X)X &= X\Lambda \\
X^*X &= I
\end{aligned}$$

Bose-Einstein condensates

- The total energy in BEC is defined as

$$E(\psi) = \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \psi(\mathbf{x})|^2 + V(\mathbf{x}) |\psi(\mathbf{x})|^2 + \frac{\beta}{2} |\psi(\mathbf{x})|^4 - \Omega \bar{\psi}(\mathbf{x}) L_z(\mathbf{x}) \right] d\mathbf{x},$$

  where $\mathbf{x} \in \mathbb{R}^d$ is the spatial coordinate vector, $\bar{\psi}$ denotes the complex conjugate of ψ, $L_z = -i(x\partial - y\partial x)$, $V(x)$ is an external trapping potential, and β, Ω are given constants.

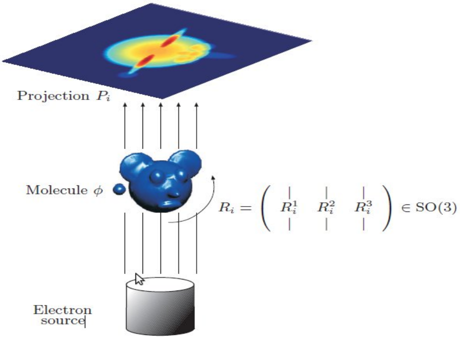
- Using a suitable discretization, we can reformulate the BEC as

$$\min_{x \in \mathbf{C}^M} \; f(x) := \frac{1}{2} x^* A x + \frac{\beta}{2} \sum_{j=1}^{M} |x_j|^4, \quad \text{s.t.} \quad \|x\|_2 = 1,$$

  where $M \in \mathcal{N}$, β is a given real constant, and $A \in \mathbf{C}^{M \times M}$ is a Hermitian matrix.

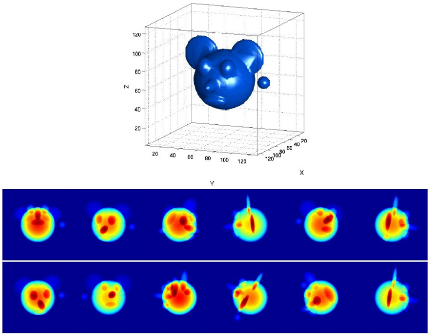
Cryo-electron microscopy reconstruction

Find 3D structure given samples of 2D images. Thanks: Amit Singer

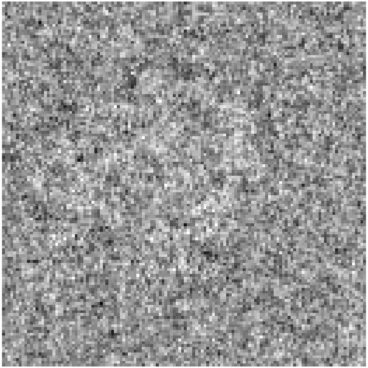


$$\min_{R_i} \quad \sum_{i=1}^{N} \|R_i c_{ij} - R_j c_{ji}\|_2^2, \quad \text{s.t.} \quad R_i^\top R_i = I_2, R_i \in \mathbb{R}^{3\times 2}$$

Challenges



Toy Example



Real Example

Both Orthogonality and Nonnegative

- $\mathcal{S}_+^{n,k} := \{X \in \mathbb{R}^{n\times k} : X^\top X = I_k,\ X \geq 0\}$
- Orthogonal NMF (ONMF): Data matrix $A \in \mathbb{R}_+^{n\times r}$, n data samples, each with r features, k clusters

$$\min_{X\in\mathcal{S}_+^{n,k}, Y\in\mathbb{R}_+^{r\times k}} \|A - XY^\top\|_F^2$$

- Orthonormal projective NMF (OPNMF) model, Yang & Oja (2010)

$$\min_{X\in\mathcal{S}_+^{n,k}} \|A - XX^\top A\|_F^2$$

- K-indicators model, Chen, Yang, Xu, Zhang & Zhang (2019)

$$\min_{X\in\mathcal{S}_+^{n,k}, Y\in\mathcal{S}^{k,k}} \|UY - X\|_F^2 \quad \text{s.t.} \quad \|X_{i,:}\|_0 = 1, i \in [n],$$

where $U \in \mathcal{S}^{n,k}$ is the features matrix extracted from the data matrix A.

Batch normalization (BN) from deep learning [1]

- Given weight vector w, the output x from the previous layer
- Batch normalization transform on $z := w^\top x$

$$BN(z) = \frac{z - \mathbf{E}[z]}{\sqrt{Var[z]}} = \frac{w^\top(x - \mathbf{E}[x])}{\sqrt{w^\top R_{xx} w}} = \frac{u^\top(x - \mathbf{E}[x])}{\sqrt{u^\top R_{xx} u}}$$

  where $u = w/\|w\|$, $\mathbf{E}[x]$ and R_{xx} are the mean and covariance of x.
- Note that $BN(w^\top x) = BN(u^\top x)$, then the wight vector satisfies

$$w \in \mathcal{G}(1, n)$$

  where $\mathcal{G}(1, n)$ is the set of 1-dimensional subspaces of $\mathbb{R}^n$.
- Deep networks with multiple layers and multiple units per layer

$$\min_{\mathcal{X} \in \mathcal{M}} \mathcal{L}(X) \text{ where } \mathcal{M} = \mathcal{G}(1, n_1) \times \cdots \times \mathcal{G}(1, n_m) \times \mathbb{R}^l$$

- dimensions of m weight vectors $n_1, \ldots, n_m$, l remaining parameters.

[1] Cho, M., Lee, J. (2017). Riemannian approach to batch normalization. In Advances in Neural Information Processing Systems (pp. 5225-5235).

Weight normalization (WN) from deep learning [2]

- Neural network: given weight matrix w, bias term b, output x from previous layer, elementwise nonlinear function ϕ

$$y = \phi(w^\top x + b),$$

- Weight normalization on w

$$\|w\|_2 = 1.$$

- Deep networks with multiple layers and multiple units per layer

$$\min_{\mathcal{X} \in \mathcal{M}} \mathcal{L}(X) \text{ where } \mathcal{M} = S^{n_1 - 1} \times \cdots \times S^{n_m - 1} \times \mathbb{R}^l$$

  where S^{n-1} is the $(n-1)$-dimensional sphere in $\mathbb{R}^n$.
- Benefits of BN and WN
  - Allow higher learning rates and train faster.
  - Make weights easier to initialize and more activation functions viable.
  - Provide a bit of regularization.
  - May give better results.

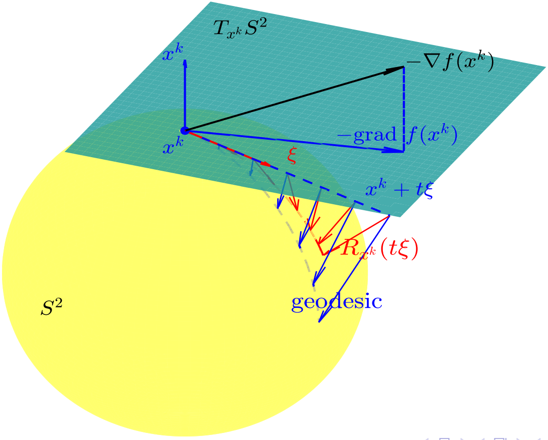
[2] Salimans, T., Kingma, D. P. (2016). Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In Advances in Neural Information

Outline

Retraction

A retraction R_x on a manifold $\mathcal{M}$ at a point x is a mapping from 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 $T_x\mathcal{M}$.
- $\mathcal{D}R_x(0_x) = \mathrm{id}_{T_x\mathcal{M}}$, where $\mathrm{id}_{T_x\mathcal{M}}$ denotes the identity mapping on $T_x\mathcal{M}$.



Curvilinear search on Riemannian manifold

Curvilinear search updating formula

$$x_{k+1} = R_{x_k}(t_k \eta_k).$$

- R_{x_k} is a retraction at x_k.
- η_k is chosen as descent direction, i.e., $\langle \mathrm{grad}\, f(x_k), \eta_k \rangle_{x_k} < 0$.
- t_k as the step size is chosen properly
  Non-monotone Armijio rule: Given $\rho, \delta \in (0, 1)$, find the smallest integer h satisfying:

  $$f(R_{x_k}(t_k \eta_k)) \leq C_k + \rho t_k \langle \mathrm{grad}\, f(x_k), \eta_k \rangle_{x_k},$$

  where $t_k = \gamma_k \delta^h$ and γ_k is the initial step size.
  $C_{k+1} = (\eta Q_k C_k + f(x_{k+1}))/Q_{k+1}$, where $C_0 = f(x_0)$,
  $Q_{k+1} = \eta Q_k + 1$ and $Q_0 = 1$.

Specialized Gradient-Type Methods

- Wen and Yin: Let $G_k = \nabla F(X_k)$, set $H = X_k G_k^\top - G_k X_k^\top$ and solve
$$Y = X + \frac{\tau}{2} H(X + Y)$$
for $Y(\tau)$. Using a step size τ, we update
$$X_{k+1} \leftarrow Y(\tau) = \left(I - \frac{\tau}{2}H\right)^{-1}\left(I + \frac{\tau}{2}H\right)X_k.$$
- Jiang and Dai: Given X_k and $D_k \in \mathcal{T}_{X_k}$,
$$\begin{aligned} W &= -(I_n - X_k X_k^\top)D_k, \; J(\tau) = I_p + \frac{\tau^2}{4}W^\top W + \frac{\tau}{2}X_k^\top D_k, \\ Y(\tau) &= (2X_k + \tau W)J(\tau)^{-1} - X_k. \end{aligned}$$
- Gao, Liu, Chen and Yuan: Given X_k and $G_k = \nabla F(X_k)$
$$V = X_k - \tau G_k, \bar{X} = (-I_n + 2V(V^\top V)^\dagger V^\top)X_k (\text{or } \mathrm{proj}_{St(n,p)}(V)),$$
$$X_{k+1} = \begin{cases} \bar{X}, & \text{if } \bar{X}^\top G_k = G_k^\top \bar{X} \\ -\bar{X}UT^\top, & o.w. \end{cases} \quad (\bar{X}^\top G = U\Lambda T^\top)$$

Classical Riemannian trust-region (RTR) method

- Absil, Baker, Gallivan: Trust-region methods on Riemannian manifold. Many other variants
- Riemannian trust-region (RTR) method:

$$\begin{cases} \min\limits_{\xi \in T_{x_k}\mathcal{M}} & m_k(\xi) := f(x_k) + \langle \mathrm{grad}\, f(x_k), \xi \rangle + \frac{1}{2} \langle \mathrm{Hess}\, f(x_k)[\xi], \xi \rangle, \\ \text{s.t.} & \|\xi\| \le \Delta_k, \end{cases}$$

  where $\mathrm{grad}\, f(x_k)$ is the Riemannian gradient and $\mathrm{Hess}\, f(x_k)$ is the Riemannian Hessian.
- Use truncated PCG to solve the subproblem
- Direct extension from Euclidean space to manifolds
- Many applications: low rank matrix completion, phase retrieval, eigenvalue computation
- Packages: Manopt, Pymanopt

Regularized Newton Method

- Our new adaptively regularized Newton (ARNT) method:

$$\begin{cases} \min & m_k(x) := \langle \nabla f(x_k), x - x_k \rangle + \frac{1}{2} \langle H_k[x - x_k], x - x_k \rangle + \frac{\sigma_k}{2} \|x - x_k\|^2, \\ \text{s.t.} & x \in \mathcal{M}, \end{cases}$$

where $\nabla f(x_k)$ and H_k are the Euclidean gradient Hessian.

- Regularized parameter update (trust-region-like strategy):
  - ratio: $\rho_k = \frac{f(z^k) - f(x^k)}{m_k(z^k)}$.
  - regularization parameter σ_k:

$$\sigma_{k+1} \in \begin{cases} (0, \sigma_k) & \text{if } \rho_k > \eta_2, & \Rightarrow \boxed{x_{k+1} = z_k} \\ [\sigma_k, \gamma_1 \sigma_k] & \text{if } \eta_1 \le \rho_k \le \eta_2, & \Rightarrow \boxed{x_{k+1} = z_k} \\ (\gamma_1 \sigma_k, \gamma_2 \sigma_k] & \text{otherwise.} & \Rightarrow \boxed{x_{k+1} = x_k} \end{cases}$$

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

Modified CG for subproblem

- Riemannian Gradient method with BB step size.
- Stiefel manifold: implicitly preserve the Lagrangian multipliers

$$\mathrm{Hess}\, m_k(x_k)[\xi] = \mathbf{P}_{x_k}(H_k[\xi] - U\mathrm{sym}((x_k)^*\nabla f(x_k))) + \tau_k \xi,$$

- Newton system for the subproblem

$$\mathrm{grad}\, m_k(x_k) + \mathrm{Hess}\, m_k(x_k)[\xi] = 0.$$

- Modified CG method

$$\xi_k = \begin{cases} s_k + \tau_k d_k & \text{if } d_k \neq 0, \\ s_k & \text{if } d_k = 0, \end{cases} \quad \text{with} \quad \tau_k := \frac{\langle d_k, \mathrm{grad}\, m_k(x_k)\rangle_{x_k}}{\langle d_k, \mathrm{Hess}\, m_k(x_k)[d_k]\rangle_{x_k}}$$

  - d_k represents and transports the negative curvature information
  - s^k corresponds to the “usual” output of the CG method.

Existing Riemannian quasi-Newton method

- Focus on the whole approximation B^k to Riemannian Hessian
$$\text{Hess}\, f(X^k) \;:\; T_{X^k}\mathcal{M} \to T_{X^k}\mathcal{M}.$$
- Riemannian BFGS method
$$B^{k+1} = \hat{B}^k - \frac{\hat{B}^k S^k((\hat{B}^k)^* S^k)^\flat}{((\hat{B}^k)^* S^k)^\flat S^k} + \frac{Y^k (Y^k)^\flat}{(Y^k)^\flat S^k}, \quad T_{X^{k+1}\mathcal{M}} \to T_{X^{k+1}\mathcal{M}}$$
where
$$\begin{aligned}
\hat{B}^k &= \mathbf{P}_{X^k}^{X^{k+1}} \circ B^k \circ (\mathbf{P}_{X^k}^{X^{k+1}})^{-1}, \text{change domain and range to } T_{X^{k+1}\mathcal{M}} \\
Y^k &= \beta_k^{-1} \text{grad}\, f(X^{k+1}) - \mathbf{P}_{X^k}^{X^{k+1}} \text{grad}\, f(X^k), \text{difference on } T_{X^{k+1}\mathcal{M}} \\
S^k &= \mathbf{P}_{X^k}^{X^{k+1}} \alpha_k \xi_k, \text{transport to } \mathrm{T}_{X^{k+1}\mathcal{M}}
\end{aligned}$$
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).

Adaptive regularized quasi-Newton method

- Riemannian Hessian of f on Stiefel manifold: $\mathrm{Hess}\, f(X)[U] = \mathbf{P}_X(\nabla^2 f(X)[U]) - U\mathrm{sym}(X^\top \nabla f(X))$
- 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.

Construction of B^k with structured f

- Assume the computational cost of $H^e(X)$ is much more expensive than that of $\mathcal{H}(X)$

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

- Quasi-Newton approximation:

$$B^k[S^k] = Y^k$$

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

- If we keep $H^c(X^k)$ and construct

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

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

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

How to choose an initial quasi-Newton approximation?

- For a linear operator A of high computational cost, the limited-memory Nyström approximation[3] $\hat{A}$ is

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

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

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

- 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 Nytröm approximation $\hat{C}_0^k$ can be utilized to further reduce the computational cost.
- More effective than the BB-type initialization (αI) in practice.

[3] 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.

Algorithms for linear eigenvalue problems

Task: Given large sparse $A = A^T \in \mathbb{R}^{n\times n}$, compute k largest eigenpairs (q_j, λ_j), $j = 1, \cdots, k$ for "large" $k \ll n$.

Our Framework:

1. A block method for subspace update (SU)
2. Augmented RR (ARR) projection

2 Block Method Variants for SU:

- Multi-power method
- Gauss Newton method

Acceleration: replace A by $\rho(A)$

Low-Rank Approximation For Eigenpair Computation

Nonlinear Least Squares:

$$X^* = \operatorname*{argmin}_{X \in \mathbb{R}^{n\times k}} \|XX^{\mathrm{T}} - A\|_{\mathrm{F}}^2.$$

GN: Large $nk \times nk$ normal equations, but with a simple structure

$$SX^{\mathrm{T}}X + XS^{\mathrm{T}}X = AX - X(X^{\mathrm{T}}X)$$

Closed-form solution for GN direction

Let $X \in \mathbb{R}^{n\times k}$ be full rank, and $\mathcal{P}_X = X(X^{\mathrm{T}}X)^{-1}X^{\mathrm{T}}$. Then

$$S(X) = (I - \mathcal{P}_X/2)\left(AX(X^{\mathrm{T}}X)^{-1} - X\right) + XC,$$

where $C^T = -C$, satisfies the normal equations. In particular, for $C = 0$,

$$S_0(X) = (I - \mathcal{P}_X/2)\left(AX(X^{\mathrm{T}}X)^{-1} - X\right)$$

is a minimum weighted-norm GN direction.

Modularity Maximization

- The modularity (*MEJ Newman, M Girvan, 2004*) is defined by

$$Q = \langle A - \frac{1}{2\lambda} dd^T, X \rangle$$

  where $\lambda = |E|$.
- The Integral modularity maximization problem:

$$\begin{array}{ll} \max & \langle A - \frac{1}{2\lambda} dd^T, X \rangle \\ \text{s.t.} & X \in \{0,1\}^{n \times n} \text{ is a partiton matrix.} \end{array}$$

- Probably hard to solve.

A Nonconvex Completely Positive Relaxation

- A nonconvex completely positive relaxation of modularity maximization:

$$\begin{aligned} &\min \langle -A + \frac{1}{2\lambda} dd^T, UU^T \rangle \\ &\text{s.t.} U \in \mathbb{R}^{n\times k} \\ &\quad \|u_i\|^2 = 1, \|u_i\|_0 \le p, i = 1, \ldots, n, \\ &\quad U \ge 0 \end{aligned}$$

- $\|u_i\|^2 = 1$: helpful in the algorithm.
- $U \ge 0$: important in theoretical proof.
- $\|u_i\|_0 \le p$: keep the sparsity.

A Nonconvex Proximal RBR Algorithm

- Let $\mathcal{U}_i := \{u_i \in \mathbb{R}^k \mid u_i \geq 0, \|u_i\|_2 = 1, \|u_i\|_0 \leq p\}$. Rewrite:

$$\min_{U \in \mathcal{U}} f(U) \equiv \langle C, UU^T \rangle, \qquad U = [u_1, u_2, \ldots, u_n]^T$$

- Proximal BCD reformulation:

$$u_i = \operatorname*{argmin}_{x \in \mathcal{U}_i} f(u_1, \ldots, u_{i-1}, x, u_{i+1}, \ldots, u_n) + \frac{\sigma}{2}\|x - \bar{u}_i\|^2$$

- Work in blocks:

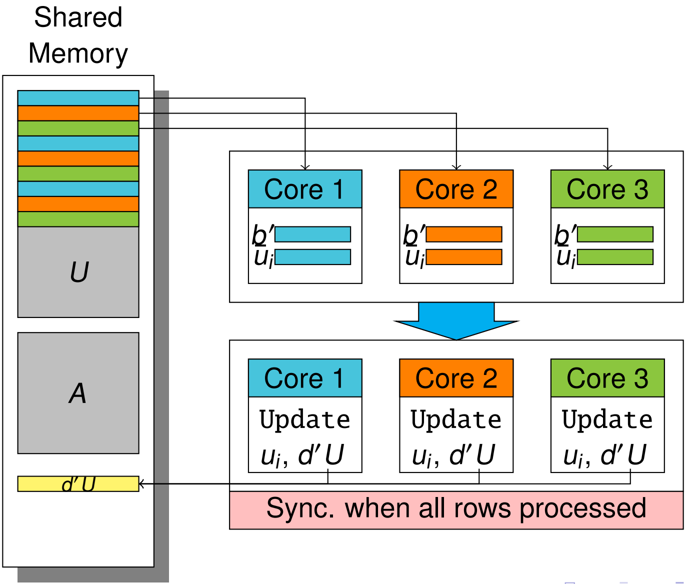
$$C = \begin{bmatrix} C_{11} & C_{1i} & C_{1n} \\ C_{i1} & c_{ii} & C_{in} \\ C_{n1} & C_{ni} & C_{nn} \end{bmatrix}, \quad UU^T = \begin{bmatrix} U_1^T U_1 & U_1^T x & U_1^T U_n \\ x^T U_1 & x^T x & x^T U_n \\ U_n^T U_1 & U_n^T x & U_n^T U_n \end{bmatrix}$$

- Note that $\|x\| = 1$. The problem is simplified to

$$u_i = \operatorname*{argmin}_{x \in \mathcal{U}_i} b^T x,$$

where $b^T = 2C^i_{-i} U_{-i} - \sigma \bar{u}_i^T$.

An Asynchronous Proximal RBR Algorithm



Optimization with nonnegative orthogonality

- Problem：

$$\min_{X \in \mathbb{R}^{n \times k}} f(X) \quad \text{s.t.} \quad X^\top X = I_k, \ X \geq 0$$

f is continuously differentiable, $\mathcal{S}_+^{n,k} := \{X \in \mathbb{R}^{n \times k} : X^\top X = I_k, \ X \geq 0\}$

- Combinatorial property: each row of X has at most one nonzero (positive) element, $\|X\|_0 \leq n$

$$X = \begin{bmatrix} \sqrt{2}/2 & 0 & 0 \\ \sqrt{2}/2 & 0 & 0 \\ 0 & \sqrt{2}/2 & 0 \\ 0 & \sqrt{2}/2 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

- Goal: find high quality orthogonal nonnegative martix

An exact penalty approach

- $\mathcal{OB}_+^{n,k} = \left\{X \in \mathbb{R}^{n\times k} : \|x_j\| = 1, x_j \geq 0, j \in [k]\right\}$
- "orth+" problem

$$\min_{X\in\mathcal{OB}_+^{n,k}} f(X) \quad \text{s.t.} \quad \|XV\|_F = 1$$

  where V can be chosen as any $V \in \mathbb{R}_{++}^{k\times r}$ $(1 \leq r \leq k)$ with $\|V\|_F = 1$
- Consider the partial penalty approach as follows:

$$\min_{X\in\mathcal{OB}_+^{n,k}} f(X) + \sigma\|XV\|_{\mathcal{F}}^2.$$

  Its global minimizer is also a global minimizer of the original problem.
- A second-order approach for solving the above problem

Outline

Convergence analysis of the SCF iteration

- Let $V := \mathcal{V}(\rho) = L^{\dagger}\rho + \mu_{xc}(\rho)^{\mathrm{T}} e$ and Hamiltonian:

$$H(V) := \frac{1}{2}L + V_{ion} + \mathrm{Diag}(V) = Q(V)\Pi(V)Q(V)^{\mathrm{T}} \quad \text{eigen-decomp}$$

- Kohn-Sham equation:

$$H(V)X(V) = X(V)\Lambda, \quad X(V)^{*}X(V) = I$$

- SCF solves a system of nonlinear equations:

$$V = \mathcal{V}(F_{\phi}(V)), \quad F_{\phi}(V) = \mathrm{diag}(X(V)X(V)^{\mathrm{T}}).$$

- Key: spectral operator $F_{\phi}(V) = \mathrm{diag}(Q(V)\phi(\Pi(V))Q(V)^{\mathrm{T}})$
- Suppose $\lambda_{p+1}(V) > \lambda_{p}(V)$. Then the directional derivative:

$$\partial_{V} F_{\phi}(V)[z] = \mathrm{diag}\left(Q(V)\left(g_{\phi}(\Pi(V)) \circ \left(Q(V)^{\mathrm{T}}\mathrm{Diag}\,(z)\,Q(V)\right)\right) Q(V)^{\mathrm{T}}\right),$$

- Rigorous convergence analysis is established

Convergence to global solutions

- Add noise to the gradient flow:

$$\mathrm{d}X(t) = -\nabla_{\mathcal{M}} F(X(t))\mathrm{d}t + \sigma(t) \circ \mathrm{d}B_{\mathcal{M}}(t),$$

  where $\mathcal{M}$ is the Stiefel manifold, and $B_{\mathcal{M}}(t)(t)$ is the Brownian motion on manifold
- One can
  - Derive and analyzed the extrinsic formulation
  - Design a numerically efficient SDE solver with strong convergence.
  - Establish overall global convergence.
  - Achieve promising numerical results in various problems.

Theorem (Convergence Results of ID)

Assuming that the local algorithm satisfies $F(X_k) \le F(X_k')$. Let the global minimum be F^*, and suppose X_{opt} to be the optimal solution obtained by ID. For any given $\epsilon > 0$ and $\zeta > 0$, $\exists \sigma > 0$, $T(\sigma) > 0$ and $N_0 > 0$ such that if $\sigma_i \le \sigma$, $T_i > T(\sigma_i)$ and $N > N_0$, $\mathbb{P}(F(X_{opt}) < F^* + \zeta) \ge 1 - \epsilon$.

Modularity minimization for community detection

- The modularity maximization problem $X = \Phi^*(\Phi^*)^\top$:

$$\begin{aligned} \max \quad & \langle A - \tfrac{1}{2\lambda} dd^T, X \rangle \\ \text{s.t.} \quad & X \in \{0,1\}^{n\times n} \text{ is a partiton matrix.} \end{aligned}$$

- Nonconvex completely positive relaxation:

$$\min_{U \in \mathbb{R}^{n\times k}} \langle -A + \frac{1}{2\lambda} dd^T, UU^T \rangle$$
$$\text{s.t.} U \geq 0, \|u_i\|^2 = 1, \|u_i\|_0 \leq p, i = 1, \dots, n$$

Theorem (Theoretical Error Bounds)

Define $G_a = \sum_{i \in C_a^*} \theta_i$, $H_a = \sum_{b=1}^k B_{ab} G_b$, $f_i = H_a \theta_i$, Under the assumption $\max_{1\leq a<b\leq k} \frac{B_{ab}+\delta}{H_a H_b} < \lambda < \min_{1\leq a\leq k} \frac{B_{aa}-\delta}{H_a^2}$ for some $\delta > 0$. Let U^* be the global optimal solution, and define $\Delta = U^*(U^*)^\top - \Phi^*(\Phi^*)^\top$. Then with high probability $\|\Delta\|_{1,\theta} \leq \frac{C_0}{\delta}\left(1 + \left(\max_{1\leq a\leq k} \frac{B_{aa}}{H_a^2} \|f\|_1\right)\right)(\sqrt{n\|f\|_1} + n)$

Analysis on a quartic-quadratic optimization problem

Definition (Model Problem)

Suppose matrix $A \in \mathbb{C}^{n\times n}$ is Hermitian and $\beta > 0$ is a constant. We consider the following minimization problem.

$$\min_{z\in\mathbb{C}^n} f(z) := \frac{1}{2} z^* A z + \frac{\beta}{2} \sum_{k=1}^{n} |z_k|^4, \quad \text{s.t. } \|z\| = 1.$$

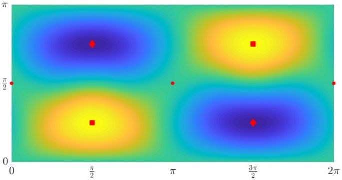
Example: Non-rotating BEC Problem

The ground state of non-rotating Bose-Einstein Condensation (BEC) problem is usually defined as the minimizer of the following dimensionless energy functional

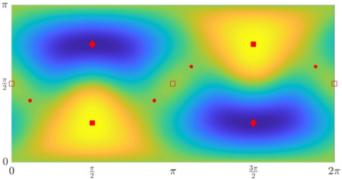
$$E(\phi) := \int_{\mathbb{R}^d} \left[\frac{1}{2} |\nabla \phi(\mathbf{x})|^2 + V(\mathbf{x}) |\phi(\mathbf{x})|^2 + \frac{\beta}{2} |\phi(\mathbf{x})|^4 \right] \mathrm{d}\,\mathbf{x},$$

where $d = 1, 2, 3$ is the dimension, $V(\mathbf{x})$ denotes the potential and $\beta \in \mathbb{R}$ is the interaction coefficient. We also need the wave function to be normalized: $\|\phi\|_{L^2(\mathbb{R}^d)} = 1$.

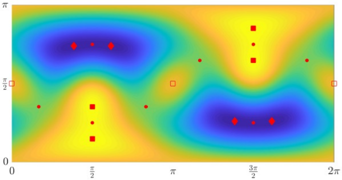
Landscape of the objective function



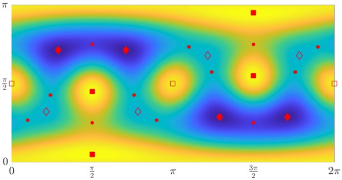
(a) $\beta = 0.25$



(b) $\beta = 0.75$



(c) $\beta = 1.25$



(d) $\beta = 3.25$

The red point marker: saddle points. Local and global minima are indicated by non-filled and filled diamond markers. The location of local and global maxima is marked by non-filled and filled squares.

Diagonal Case

Theorem (An Inequality on Perturbation)

Denote $f_\sigma(\mathbf{z}) = \frac{1}{2}\mathbf{z}^*(A + \sigma W)\mathbf{z} + \frac{\beta}{2}\|z_k\|_4^4$*, where* A *is a diagonal matrix,* W *is the Hermitian noise and* $\sigma > 0$ *is the magnitude of the noise. Suppose* $\mathbf{z}_\theta = (r_1 e^{i\theta_1}, \ldots, r_n e^{i\theta_n})^T$ *is a global minimizer of* $f_0(\mathbf{z})$ *and* $\mathbf{x} = (s_1 e^{i\phi_1}, \ldots, s_n e^{i\phi_n})^T$ *is a stationary point of* $f_\sigma(\mathbf{z})$ *that satisfies* $f_\sigma(\mathbf{x}) \le f_\sigma(\mathbf{z}_\theta)$*. Then we have*

$$\|\mathbf{x} - \mathbf{z}_\theta\|_4 \le \sqrt[3]{2\sigma\|W\|_4/\beta} \le \sqrt[3]{2\sigma\|W\|_2 n^{1/4}/\beta}.$$

Remark

Further if we have W is a Gaussian random matrix, it has been proved that $\|W\|_2 \le 3\sqrt{n}$ with probability at least $1 - 2n^{-5/4} - e^{-n/2}$. Then we know with the same probability

$$\|\mathbf{x} - \mathbf{z}_\theta\|_4 \le \sqrt[3]{6\sigma/\beta} \cdot n^{1/4}.$$

Geometric Analysis In Real Case

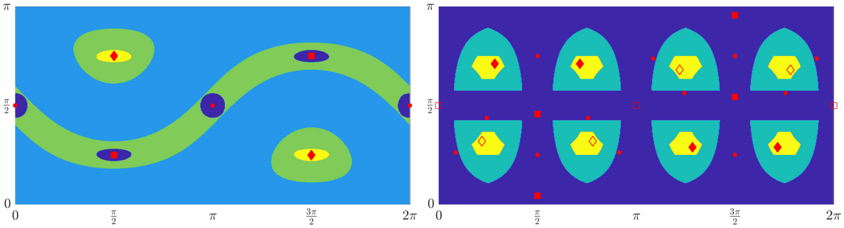
Theorem

Suppose that the coefficient β satisfies $\beta \geq \frac{8n}{n-1}(1+\gamma)\rho n^{3/2}$ for some given $\gamma > 0$. Then, the function f has the $(C_\gamma \rho, \frac{\gamma}{\sqrt{2}}\rho, C_\gamma \rho)$-strict-saddle property with $C_\gamma := \frac{4}{n-1}(1+\gamma)n^{3/2} - 1$.

Three Regions

1. (Strong convexity). $\mathcal{R}_1 = \{z \in \mathbb{S}^{n-1} : \max_{1\leq k\leq n}|z_k^2 - 1/n| \leq 1/2n\}$.
2. (Large gradient). $\mathcal{R}_2 = \{z \in \mathbb{S}^{n-1} : \max_{1\leq k\leq n}|z_k^2 - 1/n| \geq 1/2n, \min_{1\leq k\leq n} z_k^2 \geq 1/12n\}$.
3. (Negative curvature). $\mathcal{R}_3 = \{z \in \mathbb{S}^{n-1} : \min_{1\leq k\leq n} z_k^2 \leq 1/12n\}$.

Geometric Analysis In Real Case



(a) $\beta = 0.2$ (b) $\beta = 3.75$

Figure (a): The overlap of the sets $\mathcal{R}_1$–$\mathcal{R}_2$ and $\mathcal{R}_2$–$\mathcal{R}_3$ is shown in green. The set $\mathcal{R}_1$ is the union of the yellow and the two surrounding green areas, while $\mathcal{R}_2$ is the union of all green and light blue areas. The region $\mathcal{R}_3$ is the union of the dark blue sets and the enclosing green area. Figure (b): the (disjoint) yellow, turquoise, and dark blue areas directly correspond to the sets $\mathcal{R}_1$, $\mathcal{R}_2$, and $\mathcal{R}_3$, respectively. Non-filled and filled diamond markers are used for local and global minima. Local and global maxima are marked by non-filled and filled squares.

Geometric Analysis In Real Case

Corollary

If $\beta > 4\rho n^2$, the problem has at least 2^n local minima. Furthermore, if $\beta > \frac{18n^3}{n-1}\rho$, then the problem has exactly 2^n local minima

Theorem

Suppose that $\beta > \frac{18n^3}{n-1}\rho$. Then, it follows

$$f(\mathbf{y}) - \min_{\mathbf{z}\in S^{n-1}} f(\mathbf{z}) \le \frac{1}{18n} \cdot \left[\min_{\mathbf{z}\in S^{n-1}} f(\mathbf{z}) - \lambda_n(A)\right], \tag{4}$$

for all local minimizer $\mathbf{y} \in S^{n-1}$ where $\lambda_n(A)$ denotes the smallest eigenvalue of the matrix A.

Geometric Analysis In Real Case

Theorem

Suppose that the gap between the two smallest eigenvalues of the matrix A satisfies $\delta := \lambda_{n-1} - \lambda_n > 0$ and let $\gamma > 0$ be given. If $\beta \le [2(\frac{7}{3} + \gamma) + (\frac{2}{3} + \gamma)\frac{\rho}{\delta}]^{-1}\delta =: b_\gamma$, then f has the $(\gamma\beta, \gamma\beta, \gamma\beta)$-strict-saddle property.

Three Regions

1. (strong convexity) $\mathcal{R}_1 = \{z | a_n^2 \ge \frac{3\beta+\rho}{\delta+\rho}, \sum a_k^2 = 1\}$,
2. (large gradient) $\mathcal{R}_2 = \{z | \sum \lambda_k^2 a_k^2 - (\sum \lambda_k a_k^2)^2 \ge 9\beta^2, \sum a_k^2 = 1\}$,
3. (negative curvature) $\mathcal{R}_3 = \{z | a_n^2 \le \frac{\delta-5\beta}{\delta+\rho}, \sum a_k^2 = 1\}$,

where $(a_1, \ldots, a_n)^T$ are coordinates of vector z under the orthogonal basis consisting of eigenvectors of matrix A.

Under the condition of the last theorem, the optimization problem has two equivalent local minima and they are global minima.

Estimation of the Kurdyka-Łojasiewicz Exponent

- Find the largest $\theta \in (0, \frac{1}{2}]$ such that for all stationary points $\mathbf{z}$, the Łojasiewicz inequality,

$$|f(\mathbf{y}) - f(\mathbf{z})|^{1-\theta} \leq \eta_{\mathbf{z}} \|\mathrm{grad}\, f(\mathbf{y})\|, \quad \forall\, \mathbf{y} \in B(\mathbf{z}, \delta_{\mathbf{z}}) \cap \mathbb{CS}^{n-1}, \tag{5}$$

  holds with some constants $\delta_{\mathbf{z}}, \eta_{\mathbf{z}} > 0$.
- Let $A = \mathrm{diag}(\mathbf{a}) \in \mathbf{C}^{n\times n}$, $\mathbf{a} \in \mathbb{R}^n$, be a diagonal matrix. Then, the largest KL exponent is at least $\frac{1}{4}$.
- Suppose $A \in \mathbb{R}^{n\times n}$ is a symmetric matrix and $\mathbf{z}$ is a stationary point satisfying

$$H := A + 2\beta \mathrm{diag}(|\mathbf{z}|^2) - 2\lambda I \succeq 0,$$

  where $\lambda = \mathbf{z}^* \nabla_{\mathbf{z}} f(\mathbf{z}) = \frac{1}{2}\mathbf{z}^* A \mathbf{z} + \beta \|\mathbf{z}\|_4^4$. Then, the largest KL exponent of (**??**) at $\mathbf{z}$ is at least $\frac{1}{4}$.

Many Thanks For Your Attention!

- Looking for Ph.D students and Postdoc
  Competitive salary as U.S and Europe
- http://bicmr.pku.edu.cn/~wenzw
- E-mail: wenzw@pku.edu.cn
- Office phone: 86-10-62744125