Online Generalized Eigenvalue Decomposition: Primal Dual Geometry and Inverse-Free Stochastic Optimization

Zhehui Chen†, Xingguo Li◇, Lin Yang◇, Jarvis Haupt*, Tuo Zhao†

†Georgia Tech ◇Princeton University *University of Minnesota

Background

Generalized Eigenvalue Decomposition (GEV) problem [2]:

$$X^* = \underset{X\in\mathbb{R}^{d\times r}}{\operatorname{argmin}} \; -\operatorname{tr}(X^\top AX) \;\text{ s. t. }\; X^\top BX = I_r, \qquad (1)$$

where $A, B \in \mathbb{R}^{d\times d}$ and B is positive semidefinite.

GEV covers *a broad family* of problems:

- Linear Discriminant Analysis
- Canonical Correlation Analysis
- Generalized Rayleigh Quotient Problem
- Sliced Inverse Regression

Popular settings:

- Finite sum: $A = \frac{1}{n}\sum_{k=1}^n A^{(k)}$ and $B = \frac{1}{n}\sum_{k=1}^n B^{(k)}$
- Online/Stochastic: $A = \mathbb{E}A^{(k)}$ and $B = \mathbb{E}B^{(k)}$ ✓

Geometric Interpretation

Recast GEV problem (1) as an *unconstrained min-max* problem by the method of Lagrange multipliers:

$$\min_X \max_Y \mathcal{L}(X,Y) = -\operatorname{tr}(X^\top AX) + \langle Y, X^\top BX - I_r\rangle. \quad (2)$$

By KKT conditions, X and Y at a stationary point satisfy

$$\begin{cases} \nabla_X \mathcal{L}(X,Y) = 2BXY - 2AX = 0 \\ \nabla_Y \mathcal{L}(X,Y) = X^\top BX - I_r = 0 \end{cases} \Longrightarrow Y = \underbrace{X^\top AX}_{\mathcal{D}(X)}.$$

For simplicity, we denote

$$\nabla\mathcal{L} \triangleq \begin{bmatrix} \nabla_X \mathcal{L}(X,Y) \\ \nabla_Y \mathcal{L}(X,Y) \end{bmatrix} = \begin{bmatrix} 2BXY - 2AX \\ X^\top BX - I_r \end{bmatrix}.$$

Definition. *Given $\mathcal{L}(X,Y)$, a pair (X,Y) is:*

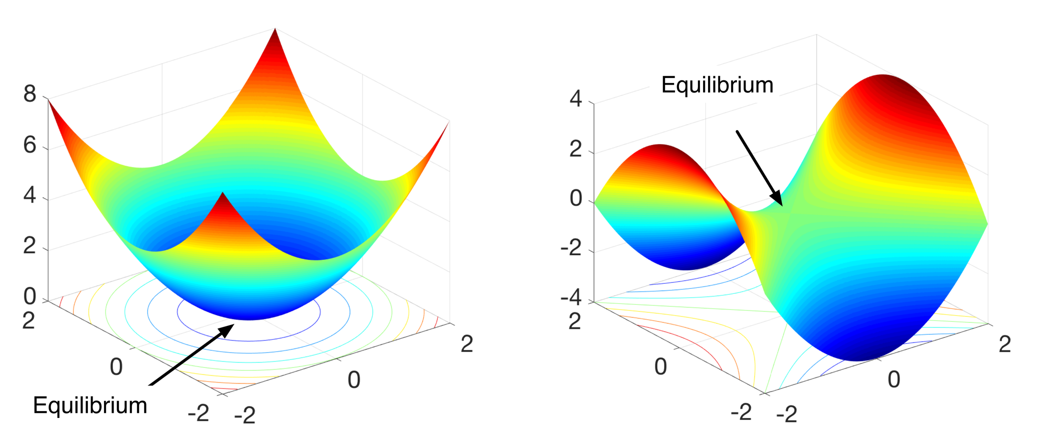
(1) *An* **equilibrium** *of $\mathcal{L}(X,Y)$, if $\nabla\mathcal{L} = 0$;*

(2) *An* **unstable equilibrium** *of $\mathcal{L}(X,Y)$, if (X,Y) is an equilibrium and for any neighborhood $\mathcal{B} \subseteq \mathbb{R}^{d\times r}$ of X, $\exists\, X_1, X_2 \in \mathcal{B}$ s.t.*

$$\mathcal{L}(X_1,Y)|_{Y=\mathcal{D}(X_1)} \le \mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)} \le \mathcal{L}(X_2,Y)|_{Y=\mathcal{D}(X_2)},$$

and $\lambda_{\min}(\nabla_X^2 \mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)}) < 0$;

(3) *A* **stable equilibrium** *of $\mathcal{L}(X,Y)$, if (X,Y) is an equilibrium, $\nabla_X^2\mathcal{L}(X,Y) \succeq 0$, and $\mathcal{L}(X,Y)$ is strongly convex over a restricted domain.*



Motivated by [1], for GEV problem (1), we aim to

- **Find** the set of **equilibria** of $\mathcal{L}(X,Y)$.
- **Distinguish stable** and **unstable equilibria**.

Invariant Group

Some terminologies:

- ***Group Action*** ϕ for a group $\mathcal{H}$ and a set $\mathcal{X}$:
  1. $\phi(\mathbf{1}, x) = x \;\; \forall x \in X$, where $\mathbf{1}$ is the identity of $\mathcal{H}$;
  2. $\phi(gh, x) = \phi(g, \phi(h,x)) \;\; \forall g, h \in H, x \in X$.
- ***Stationary Invariant Group*** of a function $f(x,y)$ w.r.t. two group actions of $\mathcal{H}$, ϕ_1 on $\mathcal{X}$ and ϕ_2 on $\mathcal{Y}$:

$$f(x,y) = f(\phi_1(g,x), \phi_2(g,y)) \;\forall x \in \mathcal{X},\; y \in \mathcal{Y},\; g \in \mathcal{H}.$$

Given $\mathcal{G} \triangleq \{\Psi \in \mathbb{R}^{r\times r} : \Psi\Psi^\top = \Psi^\top\Psi = I_r\}$, $\mathcal{L}(X,Y)$ in (2) has a stationary invariant group w.r.t two action groups of $\mathcal{G}$, ϕ_1 on $\mathbb{R}^{d\times r}$ and ϕ_2 on $\mathbb{R}^{r\times r}$:

$$\phi_1(\Psi, X) = \Psi X, \;\; \phi_2(\Psi, Y) = \Psi^\top Y\Psi.$$

$$\Longrightarrow \boxed{\mathcal{L}(X,Y) = \mathcal{L}(X\Psi, \Psi^\top Y\Psi) \;\; \forall (X,Y),\; \Psi \in \mathcal{G}.}$$

$$\Downarrow$$

- ***Equivalence Relation***: $(X,Y) \sim (X\Psi, \Psi^\top Y\Psi)$.

Symmetry Property

Notations: Symmetric Matrix M, Dimension d, Index Set $\mathcal{I}$.
Complement Index Set: $\mathcal{I}^\perp = [d]\backslash\mathcal{I}$, where $[d] = \{1, ..., d\}$;
Column Submatrix of M indexed by $\mathcal{I}$: $M_{:,\mathcal{I}}$;
Eigenvalue Decomposition of M: $M = O^M\Lambda^M(O^M)^\top$;
Index Sets with r elements: $\mathcal{X}_d^r \triangleq \{\mathcal{I} \subseteq [d], |\mathcal{I}| = r\}$.

Assumption. *Given a symmetric matrix $A \in \mathbb{R}^{d\times d}$ and a positive definite matrix $B \in \mathbb{R}^{d\times d}$, the eigenvalues of $\widetilde{A} = B^{-\frac{1}{2}}AB^{-\frac{1}{2}}$, denoted by $\lambda_1^{\widetilde{A}}, ..., \lambda_d^{\widetilde{A}}$, satisfy*

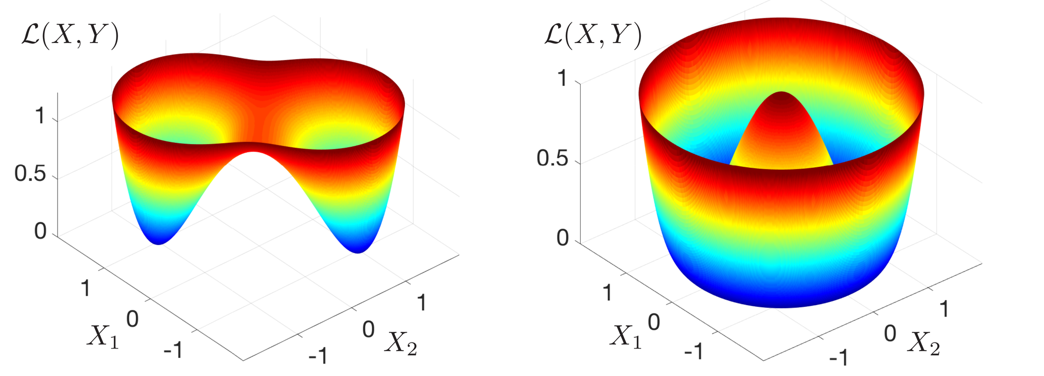
$$\lambda_1^{\widetilde{A}} \ge \cdots \ge \lambda_r^{\widetilde{A}} > \lambda_{r+1}^{\widetilde{A}} \ge \cdots \ge \lambda_d^{\widetilde{A}}.$$

Theorem 1. *Suppose Assumption holds. Then $(X, \mathcal{D}(X))$ is an equilibrium of $\mathcal{L}(X,Y)$, if and only if X can be written as*

$$X = (O^B(\Lambda^B)^{-\frac{1}{2}}O^{\widetilde{A}}_{:,\mathcal{I}}) \cdot \Psi,$$

where index set $\mathcal{I} \in \mathcal{X}_d^r$ and $\Psi \in \mathcal{G}$.

Remark. *Under the* ***equivalence relation****, there are $\binom{d}{r}$ equilibria of $\mathcal{L}(X,Y)$. Each corresponds to an $O^{\widetilde{A}}_{:,\mathcal{I}}$. Whole equilibria set is generated by $O^{\widetilde{A}}_{:,\mathcal{I}}$'s with the transformation matrix $O^B(\Lambda^B)^{-\frac{1}{2}}$ and the invariant group action induced by $\mathcal{G}$.*



Unstable Equilibria vs. Stable Equilibria

Denote the **Hessian matrix** of $\mathcal{L}(X,Y)$ w.r.t. X as

$$H_X \triangleq \nabla_X^2 \mathcal{L}(X,Y)|_{Y=\mathcal{D}(X)}.$$

Theorem 2. *Suppose Assumption holds, and $(X, \mathcal{D}(X))$ is an equilibrium in (2). By Theorem 1, X can be represented as $X = (O^B(\Lambda^B)^{-\frac{1}{2}}O^{\widetilde{A}}_{:,\mathcal{I}}) \cdot \Psi$ for some $\Psi \in \mathcal{G}$ and $\mathcal{I} \in \mathcal{X}_d^r$. Then, if $\mathcal{I} \neq [r]$, $(X, \mathcal{D}(X))$ is unstable with*

$$\lambda_{\min}(H_X) \le \frac{2(\lambda^{\widetilde{A}}_{\max\,\mathcal{I}} - \lambda^{\widetilde{A}}_{\min\,\mathcal{I}^\perp})}{\|X_{:,\min\,\mathcal{I}^\perp}\|_2^2} < 0,$$

where $\lambda^{\widetilde{A}}_{\max\,/\,\min\,\mathcal{I}} = \max/\min_{i\in\mathcal{I}} \lambda_i^{\widetilde{A}}$, and $\lambda_i^{\widetilde{A}}$ is the i-th leading eigenvalue of $\widetilde{A}$;
Otherwise, we have $H_X \succeq 0$ and rank$(H_X) = dr - r(r-1)/2$. Moreover, $(X, \mathcal{D}(X))$ is a stable equilibrium of problem (2).

Remark. *When $\mathcal{I} = [r]$, all directions in the null space of H_X, i.e.,* **non-increasing directions***, essentially point to the primal variables of* **other stable equilibria***; When $\mathcal{I} \neq [r]$, due to the* **negative curvature***, these equilibria are unstable.*

Inverse-Free Stochastic Optimization

Our stochastic GHA (SGHA) algorithm is a **primal-dual** stochastic optimization algorithm in nature. Given $A^{(k)}, B^{(k)} \in \mathbb{R}^{d\times d}$ that are independently sampled from the distribution associated with A and B at k-th iteration, SGHA updates the primal variable by

$$X^{(k+1)} \leftarrow X^{(k)} - \eta \underbrace{\left(B^{(k)}X^{(k)}Y^{(k)} - A^{(k)}X^{(k)}\right)}_{\text{Stochastic Approximation of } \nabla_{\mathbf{X}}\mathcal{L}(\mathbf{X}^{(k)}, \mathbf{Y}^{(k)})}, \quad (3)$$

where $\eta > 0$ is learning rate. Then it updates dual variable as

$$Y^{(k+1)} \leftarrow \underbrace{X^{(k)\top}A^{(k)}X^{(k)}}_{\text{Stochastic Approximation of } \mathbf{X}^{(k)\top}\mathbf{A}\mathbf{X}^{(k)}}, \quad (4)$$

Combining (3) and (4), we have a **dual-free** update as

$$X^{(k+1)} \leftarrow X^{(k)} - \eta\left(B^{(k)}X^{(k)}X^{(k)\top} - I_d\right)A^{(k)}X^{(k)}.$$

- Simple and **easy to implement**.
- **No matrix inversion** in each iteration.
- Only need **simple initial** (one random vector with each entry independently following a mean zero and variance $\frac{1}{d}$ normal distribution).

Acknowledgment

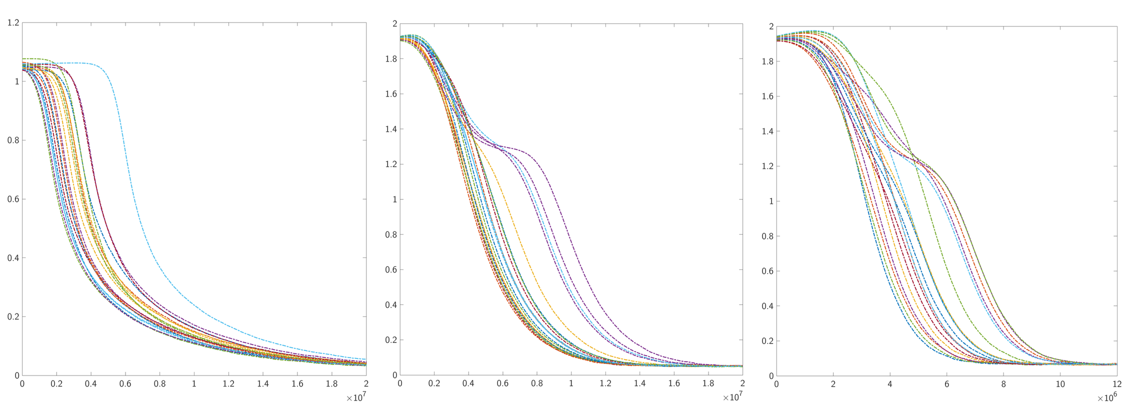
Georgia Tech · Princeton University · University of Minnesota

Numerical Results

Let $O(d)$ denote the d by d orthogonal matrices group. We have three different experiments. Their settings are as follows:

Setting 1 rank 1; **deterministic**;	$\eta = 1e-4$; $A_{ii} = \frac{1}{100}, \;\; \forall i \in [d]$; $A_{ij} = 0.5/100, \;\; \forall i \neq j$; $B_{ij} = 0.5^{\|i-j\|}/3, \;\; \forall i \neq j$.
Setting 2 rank 3; **random**; A, B **convertible**;	$\eta = 5e-5$; random $U \in O(d)$; $A = U \cdot \operatorname{diag}(1,1,1,0.1,...,0.1) \cdot U^\top$; $B = U \cdot \operatorname{diag}(2,2,2,1,...,1) \cdot U^\top$.
Setting 3 rank 3; **random**; A, B **unconvertible**;	$\eta = 2.5e-5$; random $U, V \in O(d)$; $A = U \cdot \operatorname{diag}(1,1,1,0.1,...,0.1) \cdot U^\top$; $B = V \cdot \operatorname{diag}(2,2,2,1,...,1) \cdot V^\top$.

In each iteration **independently sample 40** random vectors from $N(0,A)$ and $N(0,B)$. Use their covariance matrices as approximations of A and B to use SGHA. **Repeat 20times**.



Horizontal axis corresponds to the number of iterations, and vertical axis corresponds to the optimization error $\|B^{1/2}X^{(t)}X^{(t)\top}B^{1/2} - B^{1/2}X^*X^{*\top}B^{1/2}\|_{\text{F}}$.
Experiments indicate SGHA converges to a global optimum.

Convergence Analysis

Assumption. *$A^{(k)}$'s and $B^{(k)}$'s are independently sampled from two different distributions $\mathcal{D}_A$ and $\mathcal{D}_B$ respectively.*
(a) All the sample's are unbiased, i.e.,

$$\mathbb{E}A^{(k)} = A, \quad \mathbb{E}B^{(k)} = B.$$

Moreover, $B \succ 0$.
(b) A and B are simultaneously orthogonal diagonalizable, i.e., there exists an orthonormal matrix O such that

$$A = O\Lambda^A O^\top \quad \text{and} \quad B = O\Lambda^B O^\top,$$

where $\Lambda^A = \operatorname{diag}(\lambda_1, ..., \lambda_d)$, $\Lambda^B = \operatorname{diag}(\mu_1, ..., \mu_d)$, $\lambda_j \neq 0$, $\forall j \in [d]$. Moreover $\frac{\lambda_1}{\mu_1} > \frac{\lambda_2}{\mu_2} \ge \cdots \ge \frac{\lambda_d}{\mu_d}$ and $\mu_{\max} = \max\{\mu_2, ..., \mu_d\}$.
(c) $A^{(k)}$ and $B^{(k)}$ satisfy the following moment conditions:

$$\mathbb{E}\|A^{(k)}\|_2^2 \le C_0, \quad \mathbb{E}\|B^{(k)}\|_2^2 \le C_1,$$

where $\|\cdot\|_2$ is the spectral norm, and C_0, C_1 are constants.

Theorem. *Suppose that Assumption holds. Given a sufficiently small pre-specified error $\epsilon > 0$, we choose a step size as*

$$\eta \asymp \frac{\epsilon \cdot \text{gap}}{d \cdot \left(\frac{1}{\mu_1}C_0 \cdot C_1 + \mu_{\max}C_1\right)},$$

where gap $= \frac{\lambda_1}{\mu_1} - \frac{\lambda_2}{\mu_2}$. Then with probability at least $\frac{3}{4}$, the number of iterations required to achieve $\|W^{(N)} - W^*\|_2^2 \le \epsilon$ is at most

$$N = \mathcal{O}\left[\frac{d\left(\mu_1^{-1} + \mu_{\max}\right)}{\epsilon \cdot \text{gap}^2 \cdot \mu_{\min}} \log\left(\frac{d^{1+\mu_{\max}/\mu_1}}{\epsilon \cdot \text{gap}}\right)\right]. \quad (5)$$

Proof Sketch:

(1) Given a random initial, the **trajectory of algorithm** can be approximated by an **ordinary differential equation (ODE)**;

(2) The **norm** of each iterate **converges to a constant**;

(3) By proper **rescaling**, the algorithm can be characterized by a **stochastic differential equation (SDE)**;

(4) Obtain the **convergence rate** by the **solution of SDE**.

References

[1] X. Li, Z. Wang, J. Lu, R. Arora, J. Haupt, H. Liu, and T. Zhao. Symmetry, saddle points, and global geometry of nonconvex matrix factorization. *arXiv preprint arXiv:1612.09296*, 2016.

[2] J. H. Wilkinson and J. H. Wilkinson. *The algebraic eigenvalue problem*, volume 87. Clarendon Press Oxford, 1965.