\[
\min_{x \in X} f(x)
\]

\[X \neq \mathbb{R}^d\]
Geometry is omnipresent
Geometry is omnipresent

Vector spaces
(so far what we saw in the course)
Geometry is omnipresent

- Vector spaces
  (so far what we saw in the course)

- Convex sets
  (probability simplex, semidefinite cone, polyhedra)
Geometry is omnipresent

- Vector spaces
  (so far what we saw in the course)

- Convex sets
  (probability simplex, semidefinite cone, polyhedra)

- Manifolds, Symm. Spaces
  (sphere, orthogonal matrices, low-rank matrices, PSD)
Geometry is omnipresent

- **Vector spaces**
  (so far what we saw in the course)

- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)

- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)
Geometry is omnipresent

- Vector spaces  
  (so far what we saw in the course)

- Convex sets  
  (probability simplex, semidefinite cone, polyhedra)

- Manifolds, Symm. Spaces  
  (sphere, orthogonal matrices, low-rank matrices, PSD)

- Metric spaces  
  (tree space, Wasserstein spaces, space-of-spaces)

- Singular manifolds  
  (pseudomanifolds, intersecting manifolds, “holes”)
Vector spaces
(so far what we saw in the course)

Convex sets
(probability simplex, semidefinite cone, polyhedra)

Manifolds, Symm. Spaces
(sphere, orthogonal matrices, low-rank matrices, PSD)

Metric spaces
(tree space, Wasserstein spaces, space-of-spaces)

Singular manifolds
(pseudomanifolds, intersecting manifolds, “holes”)

Geometry is omnipresent
Geometry is omnipresent

- **Vector spaces**
  (so far what we saw in the course)

- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)

- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

- **Singular manifolds**
  (pseudomanifolds, intersecting manifolds, “holes”)

**Aim:** Geometry for foundational theory, algorithms, enable applications
Geometry is omnipresent

- **Vector spaces**
  (so far what we saw in the course)

- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)

- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

- **Singular manifolds**
  (pseudomanifolds, intersecting manifolds, “holes”)

**Aim:** Geometry for foundational theory, algorithms, enable applications

---

Machine Learning
Graphics
Robotics
Control
Computer Vision
Chip Design
NLP
Statistics
Networks
Biology
Health
Reinf. Learning
...

---

Suvrit Sra (suvrit@mit.edu)

6.881 Optimization for Machine Learning
(4/27/21 Lecture 17)
Example: Riemannian optimization

Vector space optimization

Orthogonality constraint

Fixed-rank constraint

Positive definite constraint

... ...

Stiefel manifold

Grassmann manifold

PSD manifold

Riemannian optimization

[Udriste, 1994; Absil et al., 2009]
Classes of function in optimization

Convex

Lipschitz

Smooth

Strongly convex
Classes of function in optimization

- Convex
- Lipschitz
- Strongly convex
- Smooth
- Geodesically
The idea of geodesic convexity

Convexity

see also: [Rápcsák 1984; Udriste 1994]

Metric spaces & curvature: [Menger; Alexandrov; Busemann; Bridson, Haefliger; Gromov; Perelman]
The idea of geodesic convexity

Convexity

\[ f \left( (1 - t)x \oplus ty \right) \leq (1 - t)f(x) + tf(y) \]

see also: [Rápcsák 1984; Udriște 1994]

Metric spaces & curvature: [Menger; Alexandrov; Busemann; Bridson, Haefliger; Gromov; Perelman]
The idea of geodesic convexity

Convexity

\[ f((1 - t)x \oplus ty) \leq (1 - t)f(x) + tf(y) \]

Geodesic convexity

see also: [Rápcsák 1984; Udriste 1994]

Metric spaces & curvature: [Menger; Alexandrov; Busemann; Bridson, Haefliger; Gromov; Perelman]
The idea of geodesic convexity

Convexity

\[ (1 - t)x + ty \]

Local opt of g-convex is global opt

Geodesic convexity

\[ f((1 - t)x \oplus ty) \leq (1 - t)f(x) + tf(y) \]

see also: [Rápcsák 1984; Udriste 1994]

Metric spaces & curvature: [Menger; Alexandrov; Busemann; Bridson, Haefliger; Gromov; Perelman]
Key concepts generalize
Key concepts generalize

\[ x + v \]
Key concepts generalize

Exponential map

\[ x + v \]

\[ x \]

\[ v \]

\[ \exp_x(v) \]
Key concepts generalize

Exponential map

\[ x + v \]

\[ v \]

\[ x \]

\[ y \]

\[ y - x \]

\[ x \]

\[ \text{Exp}_x(v) \]
Key concepts generalize

Exponential map

Inverse exponential map
Key concepts generalize

Exponential map

Inverse exponential map

lengths, angles, differentiation, vector translation, etc.
First-order algorithms

$$\min_{x \in \mathcal{X} \subseteq \mathcal{M}} f(x)$$
First-order algorithms

\[
\min_{x \in \mathcal{X} \subset \mathcal{M}} f(x)
\]

Assume: we can obtain exact or stochastic gradients
First-order algorithms

\[
\min_{x \in \mathcal{X} \subset \mathcal{M}} f(x)
\]

Assume: we can obtain exact or stochastic gradients

Gradient descent
\[
x \leftarrow x - \eta \nabla f(x)
\]

GD on manifolds
\[
x \leftarrow \text{Exp}_x (-\eta \nabla f(x))
\]

Exp$_x(v)$

Exp$_x^{-1}(y)$
Can we obtain global iteration complexity bounds for first-order optimization?
Can we obtain *global iteration complexity* bounds for first-order optimization?

**Global Complexity**

- Gradient Descent
- Stochastic Gradient Descent
- Coordinate Descent
- Accelerated Gradient Descent
- Fast Incremental Gradient
- ... ...

\[ \mathbb{E}[f(x_a) - f(x^*)] \leq ? \]

**Convex Optimization**

[Nemirovski-Yudin 1983]
[Nesterov 2003]
*many more works too...*

**Manifold Optimization**

Suvrit Sra (suvrit@mit.edu) 6.881 Optimization for Machine Learning (4/27/21 Lecture 17)
Example Manifolds
# Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>$\mathbb{R}^{m \times n}, \mathbb{C}^{m \times n}$</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T }^k$</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X + X^T = 0 }^k$</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>${X \in \mathbb{R}^{m \times n} : X1_n = 0_m }$</td>
</tr>
<tr>
<td>Sphere</td>
<td>${X \in \mathbb{R}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>${X \in \mathbb{R}^{n \times n} : |X|_F = 1, X = X^T }$</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>${X \in \mathbb{C}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>${X \in \mathbb{R}^{n \times m} : |X1| = \cdots = |X_m| = 1 }$</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
### Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>$\mathbb{R}^{m \times n}, \mathbb{C}^{m \times n}$</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T }^k$</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X + X^T = 0 }^k$</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>${X \in \mathbb{R}^{m \times n} : X1_n = 0_m }$</td>
</tr>
<tr>
<td>Sphere</td>
<td>${X \in \mathbb{R}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>${X \in \mathbb{R}^{n \times n} : |X|_F = 1, X = X^T }$</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>${X \in \mathbb{C}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>${X \in \mathbb{R}^{n \times m} : |X_1| = \cdots = |X_m| = 1 }$</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
## Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>$\mathbb{R}^{m \times n}, \mathbb{C}^{m \times n}$</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T }^k$</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X + X^T = 0 }^k$</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>${X \in \mathbb{R}^{m \times n} : X1_n = 0_m }$</td>
</tr>
<tr>
<td>Sphere</td>
<td>${X \in \mathbb{R}^{n \times m} :</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>${X \in \mathbb{R}^{n \times n} :</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>${X \in \mathbb{C}^{n \times m} :</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>${X \in \mathbb{R}^{n \times m} :</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
## Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>( \mathbb{R}^{m \times n}, \mathbb{C}^{m \times n} )</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>{ X \in \mathbb{R}^{n \times n} : X = X^T }^k )</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>{ X \in \mathbb{R}^{n \times n} : X + X^T = 0 }^k )</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>{ X \in \mathbb{R}^{m \times n} : XL_n = 0_m }</td>
</tr>
<tr>
<td>Sphere</td>
<td>{ X \in \mathbb{R}^{n \times m} : |X|_F = 1 }</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>{ X \in \mathbb{R}^{n \times n} : |X|_F = 1, X = X^T }</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>{ X \in \mathbb{C}^{n \times m} : |X|_F = 1 }</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>{ X \in \mathbb{R}^{n \times m} : |X_1| = \cdots = |X_m| = 1 }</td>
</tr>
<tr>
<td>Stiefel manifold</td>
<td>{ X \in \mathbb{R}^{n \times p} : X^T X = I_p }^k )</td>
</tr>
<tr>
<td>Complex Stiefel manifold</td>
<td>{ X \in \mathbb{C}^{n \times p} : X^* X = I_p }^k )</td>
</tr>
<tr>
<td>Generalized Stiefel manifold</td>
<td>{ X \in \mathbb{R}^{n \times p} : X^T BX = I_p } for some ( B &gt; 0 )</td>
</tr>
<tr>
<td>Stiefel manifold, stacked</td>
<td>{ X \in \mathbb{R}^{md \times k} : (XX^T)_{ii} = I_d }</td>
</tr>
<tr>
<td>Grassmann manifold</td>
<td>{ \text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T X = I_p }^k</td>
</tr>
<tr>
<td>Complex Grassmann manifold</td>
<td>{ \text{span}(X) : X \in \mathbb{C}^{n \times p}, X^T X = I_p }^k</td>
</tr>
<tr>
<td>Generalized Grassmann manifold</td>
<td>{ \text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T BX = I_p } for some ( B &gt; 0 )</td>
</tr>
<tr>
<td>Rotation group</td>
<td>{ R \in \mathbb{R}^{n \times n} : R^T R = I_n, \det(R) = 1 }^k</td>
</tr>
<tr>
<td>Special Euclidean group</td>
<td>{ (R, t) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n : R^T R = I_n, \det(R) = 1 }^k</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
# Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>$\mathbb{R}^{m \times n}, \mathbb{C}^{m \times n}$</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T }^k$</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X + X^T = 0 }^k$</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>${X \in \mathbb{R}^{m \times n} : XL_n = 0_m }$</td>
</tr>
<tr>
<td>Sphere</td>
<td>${X \in \mathbb{R}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>${X \in \mathbb{R}^{n \times n} : |X|_F = 1, X = X^T }$</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>${X \in \mathbb{C}^{n \times m} : |X|_F = 1 }$</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>${X \in \mathbb{R}^{n \times m} : |X_1| = \cdots = |X_m| = 1 }$</td>
</tr>
<tr>
<td>Stiefel manifold</td>
<td>${X \in \mathbb{R}^{n \times p} : X^T X = I_p }^k$</td>
</tr>
<tr>
<td>Complex Stiefel manifold</td>
<td>${X \in \mathbb{C}^{n \times p} : X^* X = I_p }^k$</td>
</tr>
<tr>
<td>Generalized Stiefel manifold</td>
<td>${X \in \mathbb{R}^{n \times p} : X^T BX = I_p }$ for some $B &gt; 0$</td>
</tr>
<tr>
<td>Stiefel manifold, stacked</td>
<td>${X \in \mathbb{R}^{md \times k} : (XX^T)_{ii} = I_d }$</td>
</tr>
<tr>
<td>Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T X = I_p }^k$</td>
</tr>
<tr>
<td>Complex Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{C}^{n \times p}, X^T X = I_p }^k$</td>
</tr>
<tr>
<td>Generalized Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T BX = I_p }$ for some $B &gt; 0$</td>
</tr>
<tr>
<td>Rotation group</td>
<td>${R \in \mathbb{R}^{n \times n} : R^T R = I_n, \det(R) = 1 }^k$</td>
</tr>
<tr>
<td>Special Euclidean group</td>
<td>${(R, t) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n : R^T R = I_n, \det(R) = 1 }^k$</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
# Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Name</th>
<th>Set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Euclidean space (complex)</td>
<td>$\mathbb{R}^{m \times n}$, $\mathbb{C}^{m \times n}$</td>
</tr>
<tr>
<td>Symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T}^k$</td>
</tr>
<tr>
<td>Skew-symmetric matrices</td>
<td>${X \in \mathbb{R}^{n \times n} : X + X^T = 0}^k$</td>
</tr>
<tr>
<td>Centered matrices</td>
<td>${X \in \mathbb{R}^{m \times n} : X1_n = 0_m}$</td>
</tr>
<tr>
<td>Sphere</td>
<td>${X \in \mathbb{R}^{n \times m} : |X|_F = 1}$</td>
</tr>
<tr>
<td>Symmetric sphere</td>
<td>${X \in \mathbb{R}^{n \times n} : |X|_F = 1, X = X^T}$</td>
</tr>
<tr>
<td>Complex sphere</td>
<td>${X \in \mathbb{C}^{n \times m} : |X|_F = 1}$</td>
</tr>
<tr>
<td>Oblique manifold</td>
<td>${X \in \mathbb{R}^{n \times m} : |X_1| = \cdots = |X_m| = 1}$</td>
</tr>
<tr>
<td>Stiefel manifold</td>
<td>${X \in \mathbb{R}^{n \times p} : X^T X = I_p}^k$</td>
</tr>
<tr>
<td>Complex Stiefel manifold</td>
<td>${X \in \mathbb{C}^{n \times p} : X^* X = I_p}^k$</td>
</tr>
<tr>
<td>Generalized Stiefel manifold</td>
<td>${X \in \mathbb{R}^{n \times p} : X^T B X = I_p}$ for some $B &gt; 0$</td>
</tr>
<tr>
<td>Stiefel manifold, stacked</td>
<td>${X \in \mathbb{R}^{md \times k} : (X X^T)_{ii} = I_d}$</td>
</tr>
<tr>
<td>Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T X = I_p}^k$</td>
</tr>
<tr>
<td>Complex Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{C}^{n \times p}, X^T X = I_p}^k$</td>
</tr>
<tr>
<td>Generalized Grassmann manifold</td>
<td>${\text{span}(X) : X \in \mathbb{R}^{n \times p}, X^T B X = I_p}$ for some $B &gt; 0$</td>
</tr>
<tr>
<td>Rotation group</td>
<td>${R \in \mathbb{R}^{n \times n} : R^T R = I_n, \det(R) = 1}^k$</td>
</tr>
<tr>
<td>Special Euclidean group</td>
<td>${(R, t) \in \mathbb{R}^{n \times n} \times \mathbb{R}^n : R^T R = I_n, \det(R) = 1}^k$</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
### Common Riemannian manifolds

<table>
<thead>
<tr>
<th><strong>Fixed-rank</strong></th>
<th>( { X \in \mathbb{R}^{m \times n} : \text{rank}(X) = k } )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-rank tensor</strong></td>
<td>Tensors of fixed multilinear rank in Tucker format</td>
</tr>
<tr>
<td><strong>Matrices with strictly positive entries</strong></td>
<td>( { X \in \mathbb{R}^{m \times n} : X_{ij} &gt; 0 \ \forall i, j } )</td>
</tr>
<tr>
<td><strong>Symmetric, positive definite matrices</strong></td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T, X &gt; 0 }^k )</td>
</tr>
<tr>
<td><strong>Symmetric positive semidefinite, fixed-rank</strong></td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k } )</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
### Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Fixed-rank</th>
<th>( { X \in \mathbb{R}^{m \times n} : \operatorname{rank}(X) = k } )</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-rank tensor</strong></td>
<td>Tensors of fixed multilinear rank in Tucker format</td>
</tr>
<tr>
<td><strong>Matrices with strictly positive entries</strong></td>
<td>( { X \in \mathbb{R}^{m \times n} : X_{ij} &gt; 0 \ \forall i, j } )</td>
</tr>
<tr>
<td><strong>Symmetric, positive definite matrices</strong></td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T, X &gt; 0 }^k )</td>
</tr>
<tr>
<td><strong>Symmetric positive semidefinite, fixed-rank</strong></td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \operatorname{rank}(X) = k } )</td>
</tr>
</tbody>
</table>

*taken from manopt.org*
## Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed-rank</td>
<td>( { X \in \mathbb{R}^{m \times n} : \text{rank}(X) = k } )</td>
</tr>
<tr>
<td>Fixed-rank tensor</td>
<td>Tensors of fixed multilinear rank in Tucker format</td>
</tr>
<tr>
<td>Matrices with strictly positive entries</td>
<td>( { X \in \mathbb{R}^{m \times n} : X_{ij} &gt; 0 \ \forall i, j } )</td>
</tr>
<tr>
<td>Symmetric, positive definite matrices</td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T, X &gt; 0 }^k )</td>
</tr>
<tr>
<td>Symmetric positive semidefinite, fixed-rank</td>
<td>( { X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k } )</td>
</tr>
</tbody>
</table>

*taken from manopt.org*
### Common Riemannian manifolds

<table>
<thead>
<tr>
<th>Manifold Type</th>
<th>Description</th>
<th>Example Set</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixed-rank</strong></td>
<td>${X \in \mathbb{R}^{m \times n} : \text{rank}(X) = k}$</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k, \text{diag}(X) = 1}$</td>
</tr>
<tr>
<td><strong>Symmetric, positive definite matrices</strong></td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T, X &gt; 0}$</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k}$</td>
</tr>
<tr>
<td><strong>Symmetric positive semidefinite, fixed-rank</strong></td>
<td>Tensors of fixed multilinear rank in Tucker format</td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k, \text{trace}(X) = 1}$</td>
</tr>
<tr>
<td><strong>Matrices with strictly positive entries</strong></td>
<td>${X \in \mathbb{R}^{m \times n} : X_{ij} &gt; 0 \ \forall i, j}$</td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X^T 1_n = 1_n$</td>
</tr>
<tr>
<td><strong>Multinomial manifold (strict simplex elements)</strong></td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X^T 1_n = 1_n$</td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X^T 1_n = 1_n$ and $X 1_n = 1_n$</td>
</tr>
<tr>
<td><strong>Symmetric positive semidefinite, fixed-rank</strong></td>
<td>${X \in \mathbb{R}^{n \times n} : X = X^T \geq 0, \text{rank}(X) = k}$</td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X^T 1_n = 1_n$</td>
</tr>
<tr>
<td><strong>Multinomial symmetric and stochastic manifold</strong></td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X 1_n = 1_n$ and $X = X^T$</td>
<td>${X \in \mathbb{R}^{n \times n} : X_{ij} &gt; 0 \ \forall i, j}$ and $X 1_n = 1_n$ and $X = X^T$</td>
</tr>
</tbody>
</table>

[taken from manopt.org]
Examples & Applications
Eigenvector problems

Eigenvector problems

Eigenvector problems

Largest eigenvector

\[
\max_{x^T x = 1} x^T A x
\]

Eigenvector problems

Largest eigenvector

\[
\max_{x^T x = 1} x^T A x
\]

Power iteration

\[
x \leftarrow \frac{A x}{\|A x\|}
\]

Eigenvector problems

Largest eigenvector

$$\max_{x^T x = 1} x^T A x$$

Power iteration

$$x \leftarrow \frac{Ax}{\|Ax\|}$$

May be viewed as Riemannian gradient descent
(albeit under another “retraction” instead of the $\text{Exp}$-map)

Eigenvector problems

Largest eigenvector

\[ \max_{x} x^T A x \]
\[ x^T x = 1 \]

Power iteration

\[ x \leftarrow \frac{A x}{\|A x\|} \]

May be viewed as Riemannian gradient descent
\(\text {(albeit under another “retraction” instead of the } \text{Exp-map)}\)

Provides transparent reasoning for global convergence rate of iter

Ref: Projection-like retractions on matrix manifolds, Pierre-Antoine Absil, Jérôme Malick.
Stochastic eigenvectors (large-scale)
Stochastic eigenvectors (large-scale)

\[ \min_{x^T x = 1} -x^T \left( \sum_{i=1}^{n} z_i z_i^T \right) x \]

\( n \) is big
Stochastic eigenvectors (large-scale)

\[
\min_{x^T x = 1} -x^T \left( \sum_{i=1}^n z_i z_i^T \right) x
\]

Lots of work on “SGD” for eigenvectors exists

[Garber, Hazan 2015; Jin, Kakade, Musco, Netrapalli, Sidford 2015; Shamir 2015, 2016]
Stochastic eigenvectors (large-scale)

Lots of work on “SGD” for eigenvectors exists

[Garber, Hazan 2015; Jin, Kakade, Musco, Netrapalli, Sidford 2015; Shamir 2015, 2016]

Simpler analysis thanks to a key geometric realization

Even though problem is geodesically non-convex, it satisfies a Riemannian Polyak-Łojasiewicz inequality

Running Riemannian SGD will obtain global optimum

[Zhang, Reddi, Sra, NIPS 2016]
### Stochastic eigenvectors (large-scale)

Lots of work on “SGD” for eigenvectors exists

[Garber, Hazan 2015; Jin, Kakade, Musco, Netrapalli, Sidford 2015; Shamir 2015, 2016]

#### Simpler analysis thanks to a key geometric realization

**Theorem 4.** Suppose $A$ has eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ and $\delta = \lambda_1 - \lambda_2$. With probability $1 - p$, the random initialization $x^0$ falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is $O\left(\frac{d}{p^2 \delta}\right)$-gradient dominated.

Running Riemannian SGD will obtain global optimum

[Zhang, Reddi, Sra, NIPS 2016]
Stochastic eigenvectors (large-scale)
Stochastic eigenvectors (large-scale)

\[
\begin{align*}
&\min_{x^T x = 1} -x^T \left( \sum_{i=1}^{n} z_i z_i^T \right) x \\
&n \text{ is big}
\end{align*}
\]
Stochastic eigenvectors (large-scale)

\[
\min_{x^T x = 1} -x^T \left( \sum_{i=1}^{n} z_i z_i^T \right) x
\]

Theorem 4. Suppose \( A \) has eigenvalues \( \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d \) and \( \delta = \lambda_1 - \lambda_2 \). With probability \( 1 - p \), the random initialization \( x^0 \) falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is \( O(\frac{d}{p^2 \delta}) \)-gradient dominated.
Stochastic eigenvectors (large-scale)

Theorem 4. Suppose $A$ has eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ and $\delta = \lambda_1 - \lambda_2$. With probability $1 - p$, the random initialization $x^0$ falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is $O\left(\frac{d}{p^2\delta}\right)$-gradient dominated.

1. A more careful initialization should improve the bound
Stochastic eigenvectors (large-scale)

\[
\min_{x^T x = 1} -x^T \left( \sum_{i=1}^{n} z_i z_i^T \right) x
\]

\[\text{n is big}\]

**Theorem 4.** Suppose $A$ has eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ and $\delta = \lambda_1 - \lambda_2$. With probability $1 - p$, the random initialization $x^0$ falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is $O\left( \frac{d}{\sqrt{\rho^2 \delta}} \right)$-gradient dominated.

1. A more careful initialization should improve the bound
2. Can we accelerate to $\sqrt{\delta}$ ?
Matrix Factorization

\[
\min_{\hat{X} \in \mathbb{R}^{m \times n}} \text{rank} \hat{X}, \quad \text{such that} \quad \hat{X}_{ij} = X_{ij} \quad \forall (i, j) \in \Omega.
\]
Matrix Factorization

\[
\min_{\hat{X} \in \mathbb{R}^{m \times n}} \text{rank} \hat{X}, \quad \text{such that} \quad \hat{X}_{ij} = X_{ij} \quad \forall (i, j) \in \Omega.
\]

\[
\min_{U \in \mathbb{R}^{m \times r}} \min_{W \in \mathbb{R}^{r \times n}} \sum_{(i,j) \in \Omega} ((UW)_{ij} - X_{ij})^2.
\]
Matrix Factorization

\[
\min_{\hat{X} \in \mathbb{R}^{m \times n}} \text{rank} \hat{X}, \quad \text{such that} \quad \hat{X}_{ij} = X_{ij} \quad \forall (i, j) \in \Omega.
\]

\[
\min_{U \in \mathbb{R}^{m \times r}} \min_{W \in \mathbb{R}^{r \times n}} \sum_{(i, j) \in \Omega} ((UW)_{ij} - X_{ij})^2.
\]

Low-rank matrix completion via preconditioned optimization on the Grassmann manifold

Nicolas Boumal\textsuperscript{a,*}, P.-A. Absil\textsuperscript{b}
Matrix Factorization

$$\min_{\hat{X} \in \mathbb{R}^{m \times n}} \text{rank } \hat{X}, \text{ such that } \hat{X}_{ij} = X_{ij} \quad \forall (i, j) \in \Omega.$$ 

$$\min_{U \in \mathbb{R}^{m \times r}} \min_{W \in \mathbb{R}^{r \times n}} \sum_{(i,j) \in \Omega} ((UW)_{ij} - X_{ij})^2.$$ 

Low-rank matrix completion via preconditioned optimization on the Grassmann manifold

Nicolas Boumal\(^a,^*\), P.-A. Absil\(^b\)

---

Riemannian Perspective on Matrix Factorization

Kwangjun Ahn\(^*1\) and Felipe Suarez\(^{†2}\)
Example: $\log(1+x)$ concave in the usual sense, but geodesically convex since $f(x^{1-t}y^t) \leq (1-t)f(x) + tf(y)$
**G-convexity for positive definite matrices**

\[ f(x) = \log(1 + x) \]

**Example:** \( \log(1+x) \) concave in the usual sense, but geodesically convex since \( f(x^{1-t}y^t) \leq (1-t)f(x) + tf(y) \)

**Geodesic from \( X \) to \( Y \)**

\[ \gamma(t) \equiv (1 - t)X \oplus tY := X^{\frac{1}{2}} (X^{-\frac{1}{2}} Y X^{-\frac{1}{2}})^t X^{\frac{1}{2}} \]

\[ f((1 - t)X \oplus tY) \leq (1 - t)f(X) + tf(Y) \]

Since \( XY \neq YX \), cannot simply use \( X^{1-t}Y^t \) as for scalars
Examples from SDP, LMI

Condition number

\[ \kappa(X) = \frac{\lambda_{\max}(X)}{\lambda_{\min}(X)} \]

Euclidean quasiconvex but log-g-convex

Generalized eigenvalue!

\[ \lambda_{\max}(A, B) = \lambda_{\max}(A^{-1}B) \]

Euclidean quasiconvex

[Boyd, Ghaoui 1993; Nesterov, Nemirovski 1991]

log-g-convex

Trace of power

\[ \log \text{tr}(X^p), \quad p \in \mathbb{R} \]

and many more…

[Sra 2017]
Example: metric learning

**Metric learning:** a fundamental problem in machine learning
Example: metric learning

**Metric learning:** a fundamental problem in machine learning

If we can judge “similarity” between data points, classification becomes easy (e.g., via nearest neighbors).
Metric learning: a fundamental problem in machine learning

If we can judge “similarity” between data points, classification becomes easy (eg via nearest neighbors)
Linear metric learning

**Input:** pairwise constraints

\[ S := \{ (x_i, x_j) \mid x_i \text{ and } x_j \text{ are in the same class} \} \]
\[ D := \{ (x_i, x_j) \mid x_i \text{ and } x_j \text{ are in different classes} \} \]

**Goal:** learn Mahalanobis distance

\[
d_A(x, y) := (x - y)^T A (x - y)
\]

**Ensure:** distances between similar points are small
distances between dissimilar points are large
Linear metric learning

**Input:** pairwise constraints

\[ S := \{ (x_i, x_j) \mid x_i \text{ and } x_j \text{ are in the same class} \} \]
\[ D := \{ (x_i, x_j) \mid x_i \text{ and } x_j \text{ are in different classes} \} \]

**Goal:** learn Mahalanobis distance

\[ d_A(x, y) := (x - y)^T A (x - y) \]

**Ensure:** distances between similar points are small
distances between dissimilar points are large
Metric learning - convex formulations

**MMC**

[Xing, Jordan, Russell, Ng 2002]

\[ d_A(x, y) := (x - y)^T A (x - y) \]
**Metric learning - convex formulations**

**MMC**

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j)
\]

such that \[
\sum_{(x_i, x_j) \in D} \sqrt{d_A(x_i, x_j)} \geq 1
\]

\[
d_A(x, y) := (x - y)^T A (x - y)
\]
Metric learning - convex formulations

MMC
[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

LMNN
[Weinberger, Saul 2005]
large-margin SDP

\[
\begin{align*}
\min_{A \succeq 0} & \quad \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) \\
\text{such that} & \quad \sum_{(x_i, x_j) \in D} \sqrt{d_A(x_i, x_j)} \geq 1
\end{align*}
\]

\[
\begin{align*}
\min_{A \succeq 0} & \quad \sum_{(x_i, x_j) \in S} [(1 - \mu)d_A(x_i, x_j) + \mu \sum_l (1 - y_{il}) \xi_{ijl}] \\
& \quad d_A(x_i, x_l) - d_A(x_i, x_j) \geq 1 - \xi_{ijl} \\
& \quad \xi_{ijl} \geq 0
\end{align*}
\]

\[d_A(x, y) := (x - y)^T A (x - y)\]
Metric learning - convex formulations

**MMC**  
[Xing, Jordan, Russell, Ng 2002]  
Semidef. Programming (SDP)

**LMNN**  
[Weinberger, Saul 2005]  
large-margin SDP

**ITML**  
[Davis, Kulis, Jain, Sra, Dhillon 2007]  
relative entropy b/w Gaussians

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j)
\]

such that

\[
\sum_{(x_i, x_j) \in D} \sqrt{d_A(x_i, x_j)} \geq 1
\]

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} \left[ (1 - \mu) d_A(x_i, x_j) + \mu \sum_l (1 - y_{il}) \xi_{ijl} \right]
\]

\[
d_A(x_i, x_l) - d_A(x_i, x_j) \geq 1 - \xi_{ijl}
\]

\[
\xi_{ijl} \geq 0
\]

\[
\min_{A \succeq 0} D_{ld}(A, A_0)
\]

such that

\[
d_A(x, y) \leq u, \quad (x, y) \in S,
\]

\[
d_A(x, y) \geq l, \quad (x, y) \in D
\]

\[
D_{ld}(A, A_0) := \text{tr}(A A_0^{-1}) - \log \det(A A_0^{-1}) - d
\]

\[
d_A(x, y) := (x - y)^T A(x - y)
\]
Metric learning - convex formulations

MMC
[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

LMNN
[Weinberger, Saul 2005]
large-margin SDP

ITML
[Davis, Kulis, Jain, Sra, Dhillon 2007]
relative entropy b/w Gaussians

Tons of other works

\[
\begin{align*}
\min_{A \succeq 0} & \quad \sum_{(x_i,x_j) \in S} d_A(x_i, x_j) \\
\text{such that} & \quad \sum_{(x_i,x_j) \in D} \sqrt{d_A(x_i, x_j)} \geq 1
\end{align*}
\]

\[
\begin{align*}
\min_{A \succeq 0} & \quad \sum_{(x_i,x_j) \in S} \left[ (1 - \mu)d_A(x_i, x_j) + \mu \sum_l (1 - y_{il})\xi_{ijl} \right] \\
& \quad d_A(x_i, x_l) - d_A(x_i, x_j) \geq 1 - \xi_{ijl} \\
& \quad \xi_{ijl} \geq 0
\end{align*}
\]

\[
\begin{align*}
\min_{A \succeq 0} & \quad D_{ld}(A, A_0) \\
& \quad \text{such that} \quad d_A(x, y) \leq u, \quad (x, y) \in S, \\
& \quad d_A(x, y) \geq l, \quad (x, y) \in D \\
D_{ld}(A, A_0) := & \quad \text{tr}(AA_0^{-1}) - \log \det(AA_0^{-1}) - d
\end{align*}
\]

\[
d_A(x, y) := (x - y)^T A (x - y)
\]

Google Scholar
"metric learning"

Articles
About 16,500 results (0.06 sec)
A new geometric approach

Euclidean idea

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) - \lambda \sum_{(x_i, x_j) \in D} d_A(x_i, x_j)
\]

\[
d_A(x, y) := (x - y)^T A (x - y)
\]
A new geometric approach

Euclidean idea

\[ d_A(x, y) := (x - y)^T A(x - y) \]

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) - \lambda \sum_{(x_i, x_j) \in D} d_A(x_i, x_j)
\]
A new geometric approach

Euclidean idea

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) - \lambda \sum_{(x_i, x_j) \in D} d_A(x_i, x_j)
\]

New idea

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) + \sum_{(x_i, x_j) \in D} d_A^{-1}(x_i, x_j)
\]

\[d_A(x, y) := (x - y)^T A(x - y)\]
A new geometric approach

Euclidean idea

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) - \lambda \sum_{(x_i, x_j) \in D} d_A(x_i, x_j)
\]

New idea

\[
\min_{A \succeq 0} \sum_{(x_i, x_j) \in S} d_A(x_i, x_j) + \sum_{(x_i, x_j) \in D} d_A^{-1}(x_i, x_j)
\]

Intuitively: If \( a > b \), then \( a^{-1} < b^{-1} \)
Geometric approach to metric learning

Collect similar points into $S$ and dissimilar into $D$

\[
S := \sum_{(x_i, x_j) \in S} (x_i - x_j)(x_i - x_j)^T,
\]

\[
D := \sum_{(x_i, x_j) \in D} (x_i - x_j)(x_i - x_j)^T
\]

scatter matrices

[Habibzadeh, Hosseini, Sra, ICML 2016]
Geometric approach to metric learning

Collect similar points into $S$ and dissimilar into $D$

$$S := \sum_{(x_i, x_j) \in S} (x_i - x_j)(x_i - x_j)^T,$$
$$D := \sum_{(x_i, x_j) \in D} (x_i - x_j)(x_i - x_j)^T$$

scatter matrices

Equivalently solve

$$\min_{A \succeq 0} h(A) := \text{tr}(AS) + \text{tr}(A^{-1}D)$$

[Habibzadeh, Hosseini, Sra, ICML 2016]
Closed form solution!

\[ \nabla h(A) = 0 \iff S - A^{-1}DA^{-1} = 0 \]
Closed form solution!

\[ \nabla h(A) = 0 \iff S - A^{-1}DA^{-1} = 0 \]

\[ A = S^{-1} \# \frac{1}{2} D \]
Geometric approach to metric learning

Closed form solution!

\[ \nabla h(A) = 0 \iff S - A^{-1}DA^{-1} = 0 \]

\[ A = S^{-1} \# \frac{1}{2} D \]

More generally

\[ \min_{A \succ 0} (1 - t)\delta^2_R(S^{-1}, A) + t\delta^2_R(D, A) \]

\[ S^{-1} \#_t D \]
Closed form solution!

\[ \nabla h(A) = 0 \iff S - A^{-1}DA^{-1} = 0 \]

\[ A = S^{-1} \#_{\frac{1}{2}} D \]

More generally

\[ \min_{A \succeq 0} (1 - t)\delta^2_R(S^{-1}, A) + t\delta^2_R(D, A) \]

Nonconvex but solvable optimally thanks to \(g\)-convexity
Experiments

Comment: May think of this as a “supervised whitening transform”

[Habibzadeh, Hosseini, Sra ICML 2016]
## Experiments

### Running time in seconds

<table>
<thead>
<tr>
<th>DATA SET</th>
<th>GMML</th>
<th>LMNN</th>
<th>ITML</th>
<th>FlatGeo</th>
</tr>
</thead>
<tbody>
<tr>
<td>SEGMENT</td>
<td>0.0054</td>
<td>77.595</td>
<td>0.511</td>
<td>63.074</td>
</tr>
<tr>
<td>LETTERS</td>
<td>0.0137</td>
<td>401.90</td>
<td>7.053</td>
<td>13543</td>
</tr>
<tr>
<td>USPS</td>
<td>0.1166</td>
<td>811.2</td>
<td>16.393</td>
<td>17424</td>
</tr>
<tr>
<td>ISOLET</td>
<td>1.4021</td>
<td>3331.9</td>
<td>1667.5</td>
<td>24855</td>
</tr>
<tr>
<td>MNIST</td>
<td>1.6795</td>
<td>1396.4</td>
<td>1739.4</td>
<td>26640</td>
</tr>
</tbody>
</table>

Comment: May think of this as a “supervised whitening transform”

[Habibzadeh, Hosseini, Sra ICML 2016]
Brascamp-Lieb Constant
Brascamp-Lieb Constant

\[
\int_{\mathbb{R}^n} \prod_{i=1}^{m} f_i(B_i x)^{p_i} \, dx \leq D^{-1/2} \prod_{i=1}^{m} \left( \int_{\mathbb{R}^n} f_i(y) \, dy \right)^{p_i}
\]
Brascamp-Lieb Constant

\[
\int_{\mathbb{R}^n} \prod_{i=1}^{m} f_i(B_ix)^{p_i} \, dx \leq D^{-1/2} \prod_{i=1}^{m} \left( \int_{\mathbb{R}^n} f_i(y) \, dy \right)^{p_i}
\]

super generalization of: sum-of-prod \( \leq \) prod-of-sum, e.g, \( \langle x, y \rangle \leq \|x\| \cdot \|y\| \)

\[
p_i > 0, f_i \geq 0 \quad \sum_{i=1}^{m} p_in_i = n
\]

powerful inequality; includes Hölder, Loomis-Whitney, Young’s, many others!

Important in: Information theory, convex geometry, probability theory
\[ \int_{\mathbb{R}^n} \prod_{i=1}^{m} f_i(B_i x)^{p_i} \, dx \leq D^{-1/2} \prod_{i=1}^{m} \left( \int_{\mathbb{R}^n} f_i(y) \, dy \right)^{p_i} \]

super generalization of: sum-of-prod \( \leq \) prod-of-sum, e.g, \( \langle x, y \rangle \leq \|x\| \cdot \|y\| \)

\[ D := \inf \left\{ \frac{\det \left( \sum_i p_i B_i^* X_i B_i \right)}{\prod_i \left( \det X_i \right)^{p_i}} \left| X_i \succ 0, n_i \times n_i \right. \right\} \]

\[ p_i > 0, f_i \geq 0 \quad \sum_{i=1}^{m} p_i n_i = n \]

powerful inequality; includes Hölder, Loomis-Whitney, Young’s, many others!

Important in: Information theory, convex geometry, probability theory
**Brascamp-Lieb constant**

\[
\min_{X_1, \ldots, X_m > 0} \log \det \left( \sum_i p_i B_i^* X_i B_i \right) - \sum_i p_i \log \det X_i
\]

- Applications to geometric complexity theory
  
  [Garg, Gurvits, Oliveira, Wigderson; Jul 2016]

- Problem has unique solution & sufficient conditions
  
  [Bennett, Carbery, Christ, Tao, 2005]

- Barthe, Carlen, Lieb, Cordero-Erasquin, McCann, …
Brascamp-Lieb constant

\[
\min_{X_1,\ldots,X_m > 0} \log \det \left( \sum_i p_i B_i^* X_i B_i \right) - \sum_i p_i \log \det X_i
\]

- Applications to geometric complexity theory
  
  [Garg, Gurvits, Oliveira, Wigderson; Jul 2016]

- Problem has unique solution & sufficient conditions
  
  [Bennett, Carbery, Christ, Tao, 2005]

- Barthe, Carlen, Lieb, Cordero-Erasquin, McCann, ...

**Prop:** This is a g-convex optimization problem
Brascamp-Lieb constant

\[
\min_{X_1, \ldots, X_m > 0} \log \det \left( \sum_i p_i B_i^* X_i B_i \right) - \sum_i p_i \log \det X_i
\]

• Applications to geometric complexity theory
  
  \[\text{[Garg, Gurvits, Oliveira, Wigderson; Jul 2016]}\]

• Problem has unique solution & sufficient conditions
  
  \[\text{[Bennett, Carbery, Christ, Tao, 2005]}\]

• Barthe, Carlen, Lieb, Cordero-Erasquin, McCann, …

**Prop:** This is a g-convex optimization problem

**Let’s look at a proof….”**
Aim: Prove $f(X^\#_t Y) \leq (1 - t)f(X) + tf(Y)$

Recall geodesic: $X^\#_t Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2}$
Proving g-convexity

**Aim:** Prove $f(X\#_t Y) \leq (1 - t)f(X) + tf(Y)$

Recall geodesic: $X\#_t Y = X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^*XB_i$ be a positive linear map (i.e., it maps psd matrices to psd matrices and is linear too)
Proving g-convexity

**Aim:** Prove \( f(X\#_t Y) \leq (1 - t)f(X) + tf(Y) \) \( \min_{X_1,\ldots,X_m \succeq 0} \log \det \left( \sum_i p_i B_i^* X_i B_i \right) - \sum_i p_i \log \det X_i \)

Recall geodesic: \( X\#_t Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2} \)

Let \( \Phi_i(X) = B_i^* X B_i \) be a positive linear map (i.e., it maps psd matrices to psd matrices and is linear too)

**Lemma A.** \( \Phi(X\#_t Y) \leq \Phi(X)\#_t \Phi(Y) \).
**Proving g-convexity**

**Aim:** Prove \( f(X\#_t Y) \leq (1 - t)f(X) + tf(Y) \)

Recall geodesic: \( X\#_t Y = X^{1/2}(X^{-1/2}YX^{-1/2})^tX^{1/2} \)

Let \( \Phi_i(X) = B_i^*XB_i \) be a positive linear map (i.e., it maps psd matrices to psd matrices and is linear too)

**Lemma A.** \( \Phi(X\#_t Y) \leq \Phi(X)\#_t\Phi(Y) \).

**Proof.** [Kubo-Ando 1980; Sra-Hosseini 2015]
**Proving g-convexity**

**Aim:** Prove \( f(X\#_t Y) \leq (1 - t)f(X) + tf(Y) \)

Recall geodesic: \( X\#_t Y = X^{1/2}(X^{-1/2}YX^{-1/2})^tX^{1/2} \)

Let \( \Phi_i(X) = B_i^*XB_i \) be a positive linear map (i.e., it maps psd matrices to psd matrices and is linear too)

**Lemma A.** \( \Phi(X\#_t Y) \leq \Phi(X)\#_t\Phi(Y) \).

**Proof.** [Kubo-Ando 1980; Sra-Hosseini 2015]

**Lemma B.** (joint concavity) \( A\#_t B + X\#_t Y \leq (A + X)\#_t(B + Y) \).
Proving g-convexity

**Aim:** Prove $f(X\#_t Y) \leq (1 - t)f(X) + tf(Y)$

Recall geodesic: $X\#_t Y = X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^*XB_i$ be a positive linear map (i.e., it maps psd matrices to psd matrices and is linear too)

**Lemma A.** $\Phi(X\#_t Y) \leq \Phi(X)\#_t \Phi(Y)$.

**Proof.** [Kubo-Ando 1980; Sra-Hosseini 2015]

**Lemma B.** (joint concavity) $A\#_t B + X\#_t Y \leq (A + X)\#_t (B + Y)$.

**Proof.** see e.g., [Bhatia 2007, “Positive definite matrices”]
Recall $X^# Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^* X B_i$ be a positive linear map
G-convexity of BL constant

Recall \( X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2} \)

Let \( \Phi_i(X) = B_i^* X B_i \) be a positive linear map

\[
\sum_i \Phi_i(X_i \# t Y_i) \leq \sum_i \Phi_i(X_i) \# t \Phi_i(Y_i) \leq \left( \sum_i \Phi_i(X_i) \right) \# t \left( \sum_i \Phi_i(Y_i) \right)
\]
Recall \( X^t Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2} \)

Let \( \Phi_i(X) = B_i^* X B_i \) be a positive linear map

\[
\sum_i \Phi_i(X_i \#_t Y_i) \leq \sum_i \Phi_i(X_i) \#_t \Phi_i(Y_i) \leq \left( \sum_i \Phi_i(X_i) \right) \#_t \left( \sum_i \Phi_i(Y_i) \right)
\]

**Lemma A.** \( \Phi(X \#_t Y) \leq \Phi(X) \#_t \Phi(Y) \).
G-convexity of BL constant

Recall $X^\#_t Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^* X B_i$ be a positive linear map

$$\sum_i \Phi_i(X_i^\#_t Y_i) \leq \sum_i \Phi_i(X_i^\#_t Y) \leq \left( \sum_i \Phi_i(X_i) \right)^\#_t \left( \sum_i \Phi_i(Y_i) \right)$$

Lemma A. $\Phi(X^\#_t Y) \leq \Phi(X)^\#_t \Phi(Y)$.

Lemma B. (joint concavity) $A^\#_t B + X^\#_t Y \leq (A + X)^\#_t (B + Y)$. 
G-convexity of BL constant

Recall $X\#_t Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^* X B_i$ be a positive linear map

$$\sum_i \Phi_i(X_i \#_t Y_i) \leq \sum_i \Phi_i(X_i) \#_t \Phi_i(Y_i) \leq \left( \sum_i \Phi_i(X_i) \right) \#_t \left( \sum_i \Phi_i(Y_i) \right)$$

**Lemma A.** $\Phi(X\#_t Y) \leq \Phi(X)\#_t \Phi(Y)$.

**Lemma B.** (joint concavity) $A\#_t B + X\#_t Y \leq (A + X)\#_t (B + Y)$.

$$\implies \log \det \left( \sum_i \Phi_i(X_i \#_t Y_i) \right)$$
G-convexity of BL constant

Recall $X_{i}^{\#t} Y = X^{1/2}(X^{-1/2} Y X^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^* X B_i$ be a positive linear map

\[ \sum_i \Phi_i(X_i^{\#t} Y_i) \leq \sum_i \Phi_i(X_i^{\#t} Y_i) \leq (\sum_i \Phi_i(X_i))^{\#t} (\sum_i \Phi_i(Y_i)) \]

**Lemma A.** $\Phi(X_{i}^{\#t} Y) \leq \Phi(X)^{\#t} \Phi(Y)$.  

**Lemma B.** (joint concavity) $A^{\#t} B + X^{\#t} Y \leq (A + X)^{\#t} (B + Y)$.

\[ \rightarrow \log \det \left( \sum_i \Phi_i(X_i^{\#t} Y_i) \right) \]
\[ \leq (1 - t) \log \det \left( \sum_i \Phi_i(X_i) \right) + t \log \det \left( \sum_i \Phi_i(Y_i) \right) \]
G-convexity of BL constant

Recall $X\#_t Y = X^{1/2}(X^{-1/2}YX^{-1/2})^t X^{1/2}$

Let $\Phi_i(X) = B_i^*X B_i$ be a positive linear map

$$\sum_i \Phi_i(X_i \#_t Y_i) \leq \sum_i \Phi_i(X_i) \#_t \Phi_i(Y_i) \leq \left( \sum_i \Phi_i(X_i) \right) \#_t \left( \sum_i \Phi_i(Y_i) \right)$$

**Lemma A.** $\Phi(X\#_t Y) \leq \Phi(X)\#_t \Phi(Y)$.

**Lemma B.** (joint concavity) $A\#_t B + X\#_t Y \leq (A + X)\#_t (B + Y)$.

$$\implies \log \det \left( \sum_i \Phi_i(X_i \#_t Y_i) \right)$$

$$\leq (1 - t) \log \det \left( \sum_i \Phi_i(X_i) \right) + t \log \det \left( \sum_i \Phi_i(Y_i) \right)$$

This is the desired geodesic convexity inequality.
An example from optimal transport

Transport mass from one place to another at lowest cost (EMD)

**Wasserstein distance**: net cost of transport, or how far is source distribution from target distribution
Wasserstein barycenters

Wasserstein distance between multivariate Gaussians

\[ d_W(X, Y) = \left[ \text{tr}(X + Y) - 2\text{tr}(X^{1/2}YX^{1/2})^{1/2} \right]^{1/2}. \]
Wasserstein barycenters

Wasserstein distance between multivariate Gaussians

\[ d_W(X, Y) = \left[ \operatorname{tr}(X + Y) - 2\operatorname{tr}(X^{1/2}YX^{1/2})^{1/2} \right]^{1/2}. \]

Wasserstein Barycenter

\[
\min_{X \succeq 0} \quad \frac{1}{N} \sum_{i=1}^{N} d_{W}^{2}(X, A_i)
\]
Wasserstein barycenters

Wasserstein distance between multivariate Gaussians

\[ d_W(X, Y) = \left[ \text{tr}(X + Y) - 2\text{tr}(X^{1/2} Y X^{1/2})^{1/2} \right]^{1/2}. \]

Wasserstein Barycenter

\[ \min_{X \succeq 0} \frac{1}{N} \sum_{i=1}^{N} d_W^2(X, A_i) \]

Actually a (Euclidean) convex optimization problem

But empirically Riemannian optimization turns out to be faster!
Recent toolboxes, tutorials

https://www.manopt.org
Recent toolboxes, tutorials

Welcome to Manopt!
A Matlab toolbox for optimization on manifolds

Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems. With Manopt, it is easy to deal with types of symmetries and constraints which arise naturally in applications, such as orthonormality and low rank.

Pymanopt

Pymanopt is a Python toolbox for optimization on manifolds, that computes gradients and Hessians automatically. It builds toolbox Manopt but is otherwise independent of it. Pymanopt aims to lower the barriers for users wishing to use state of the art for optimization on manifolds, by relying on automatic differentiation for computing gradients and Hessians, saving users time from potential calculation and implementation errors.

https://www.manopt.org

https://www.pymanopt.org
Recent toolboxes, tutorials

Welcome to Manopt!
A Matlab toolbox for optimization on manifolds

Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems. With Manopt, it is easy to incorporate types of symmetries and constraints which arise naturally in applications, such as orthonormality and low rank.

Pymanopt

Pymanopt is a Python toolbox for optimization on manifolds, that computes gradients and Hessians automatically. It builds on the Manopt toolbox but is otherwise independent of it. Pymanopt aims to lower the barriers for users wishing to use state of the art optimization methods for optimization on manifolds, by relying on automatic differentiation for computing gradients and Hessians, saving users time and effort from potential calculation and implementation errors.

See also: https://manoptjl.org/stable/
Recent toolboxes, tutorials

Welcome to Manopt!
A Matlab toolbox for optimization on manifolds
Optimization on manifolds is a powerful paradigm to address nonlinear optimization problems. With Manopt, it is easy to express types of symmetries and constraints which arise naturally in applications, such as orthonormality and low rank.

Pymanopt
Pymanopt is a Python toolbox for optimization on manifolds, that computes gradients and Hessians, akin to the Matlab toolbox Manopt but is otherwise independent of it. Pymanopt aims to lower the barriers for users of Manopt who are not Matlab users for optimization on manifolds, by relying on automatic differentiation for computing gradients and Hessians, protecting them from potential calculation and implementation errors.

See also: https://manoptjl.org/stable/

New book: