Optimization for Machine Learning

Lecture 17: Geometric Optimization — I

6.881: MIT

Suvrit Sra

Massachusetts Institute of Technology

April 27, 2021

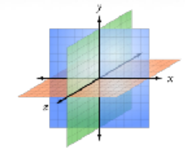
$$\min_{x \in X} \; f(x)$$

$$X \neq \mathbb{R}^d$$

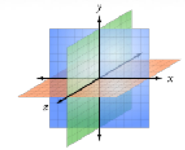
Geometry is omnipresent

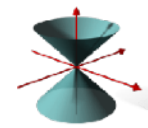
Geometry is omnipresent

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

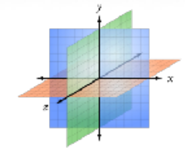


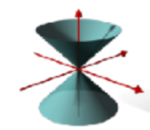
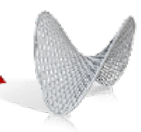
Geometry is omnipresent

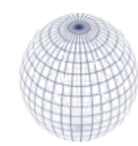
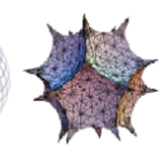
- **Vector spaces** 
  (so far what we saw in the course)

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

Geometry is omnipresent

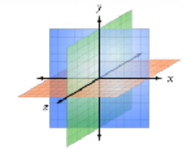
- **Vector spaces** 
  (so far what we saw in the course)

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

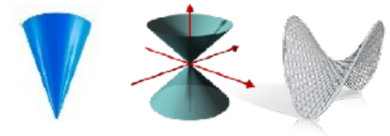
- **Manifolds, Symm. Spaces**   
  (sphere, orthogonal matrices, low-rank matrices, PSD)

Geometry is omnipresent

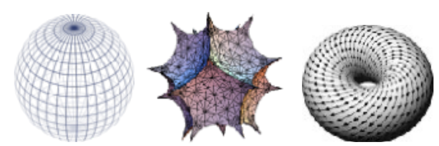
- **Vector spaces**
  (so far what we saw in the course)



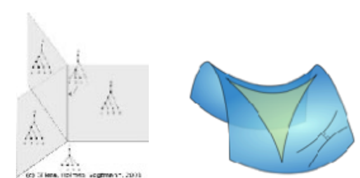
- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)



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

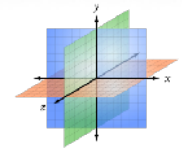


- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

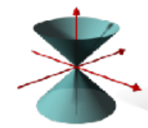


Geometry is omnipresent

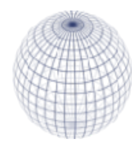
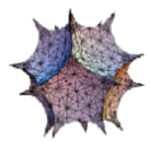
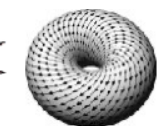
- **Vector spaces**
  (so far what we saw in the course)



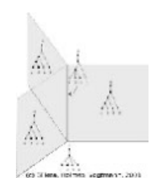
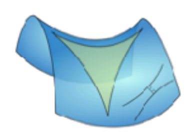
- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

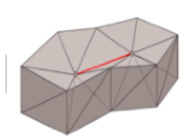
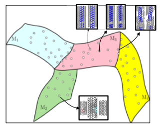

- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)

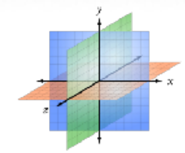
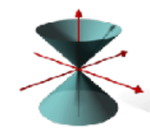
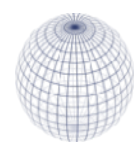
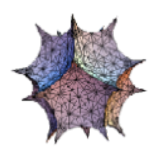
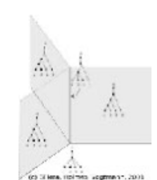
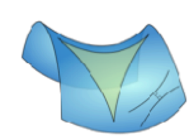
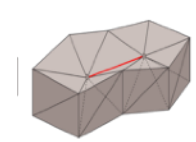
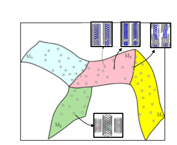
- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

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

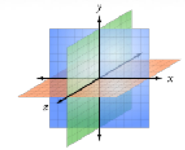
Geometry is omnipresent

- **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")

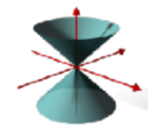
Machine Learning
Graphics
Robotics
Control
Computer Vision
Chip Design
NLP
Statistics
Networks
Biology
Health
Reinf. Learning
...

Geometry is omnipresent

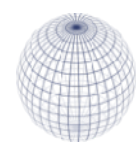
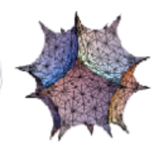
- **Vector spaces**
  (so far what we saw in the course)



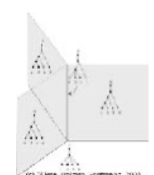
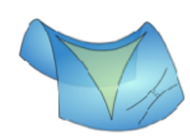
- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

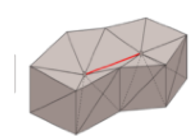
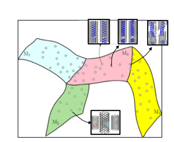
- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)

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

- **Singular manifolds**
  (pseudomanifolds, intersecting manifolds, "holes")

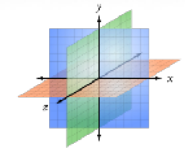
 

Machine Learning
Graphics
Robotics
Control
Computer Vision
Chip Design
NLP
Statistics
Networks
Biology
Health
Reinf. Learning
…

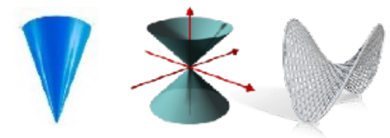
Aim: Geometry for foundational theory, algorithms, enable applications

Geometry is omnipresent

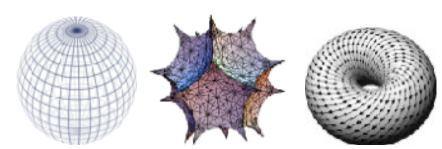
- **Vector spaces**
  (so far what we saw in the course)



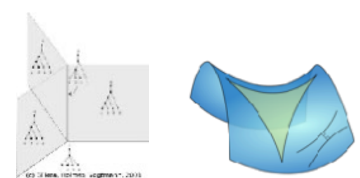
- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)



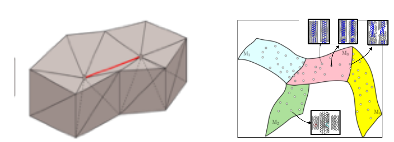
- **Manifolds, Symm. Spaces**
  (sphere, orthogonal matrices, low-rank matrices, PSD)



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



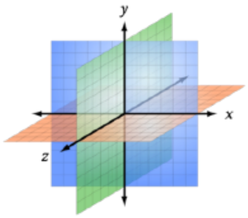
- **Singular manifolds**
  (pseudomanifolds, intersecting manifolds, "holes")



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

Aim: Geometry for foundational theory, algorithms, enable applications

Example: Riemannian optimization



Vector space optimization

Orthogonality constraint

Fixed-rank constraint

Positive definite constraint

… …

⟺

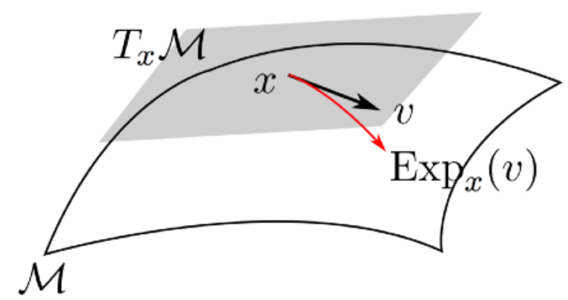
Stiefel manifold

Grassmann manifold

PSD manifold

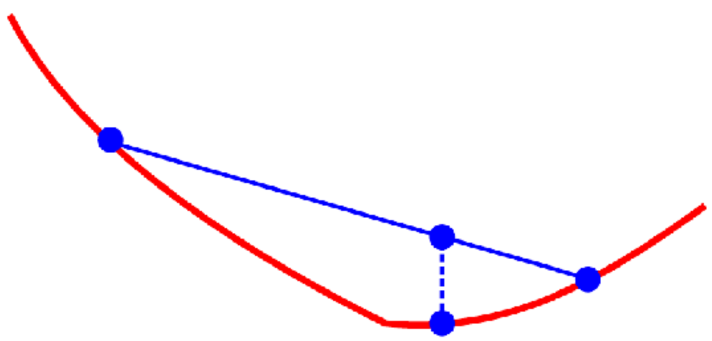
… …

Riemannian optimization

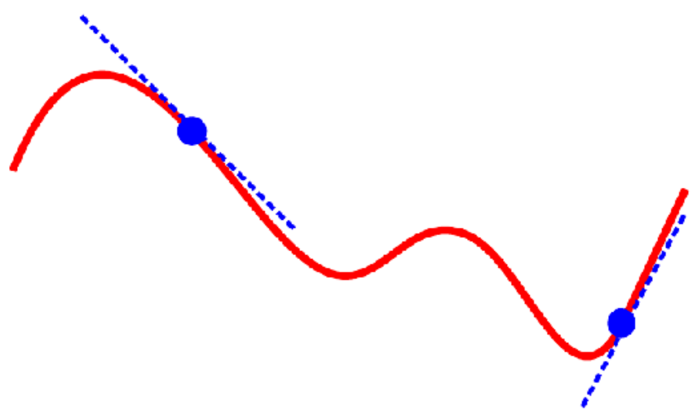


[Udriste, 1994; Absil et al., 2009]

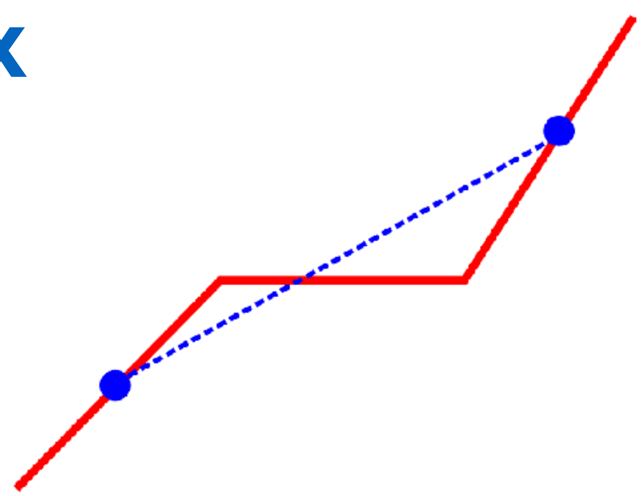
Classes of function in optimization



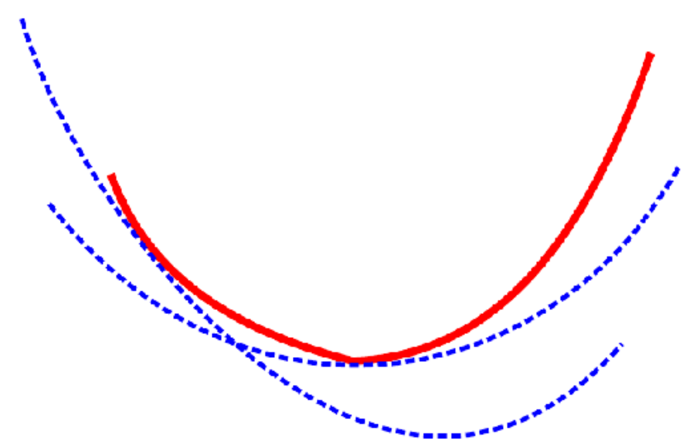
Convex



Smooth

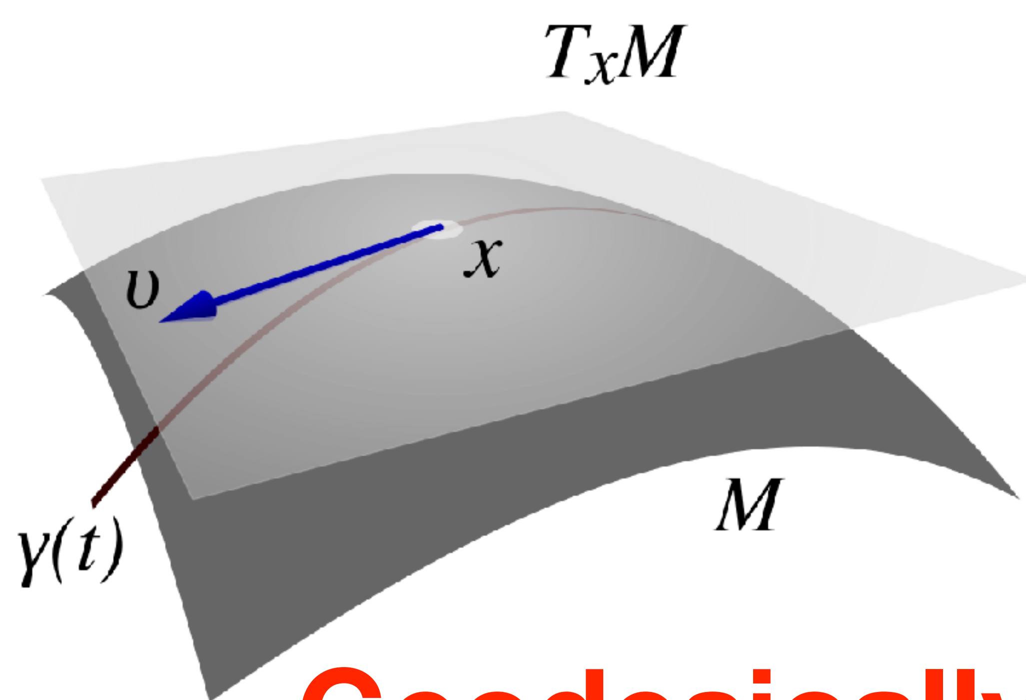


Lipschitz

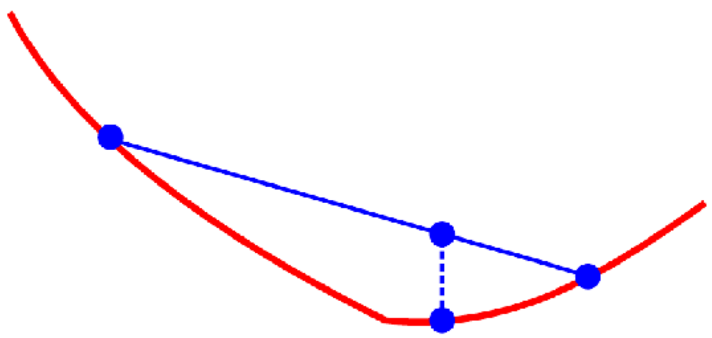


Strongly convex

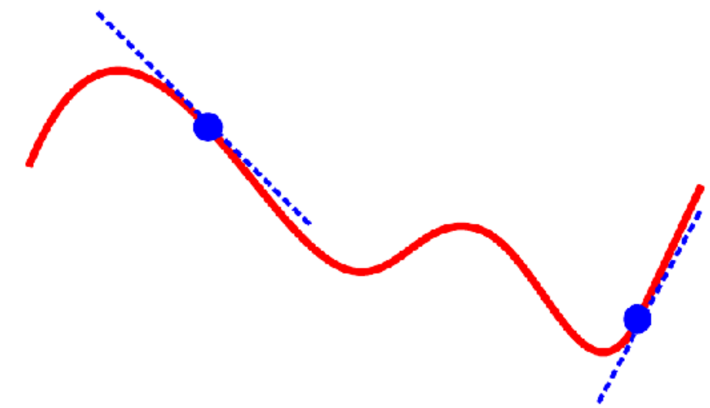
Classes of function in optimization



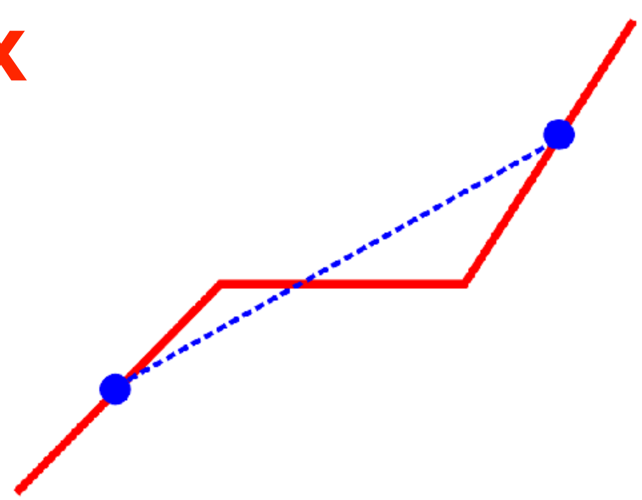
Geodesically



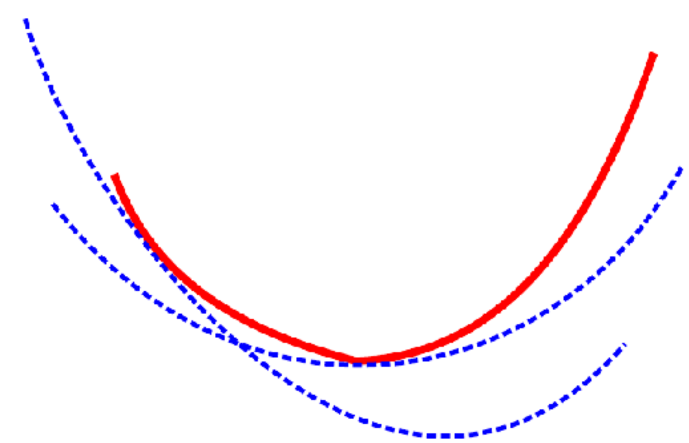
Convex



Smooth



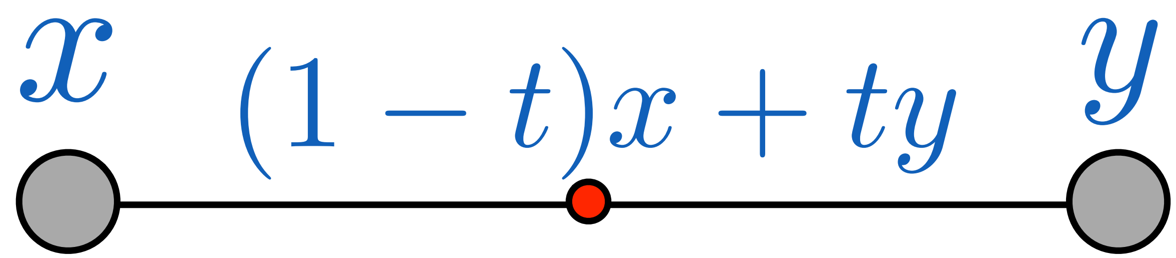
Lipschitz



Strongly convex

The idea of geodesic convexity

Convexity

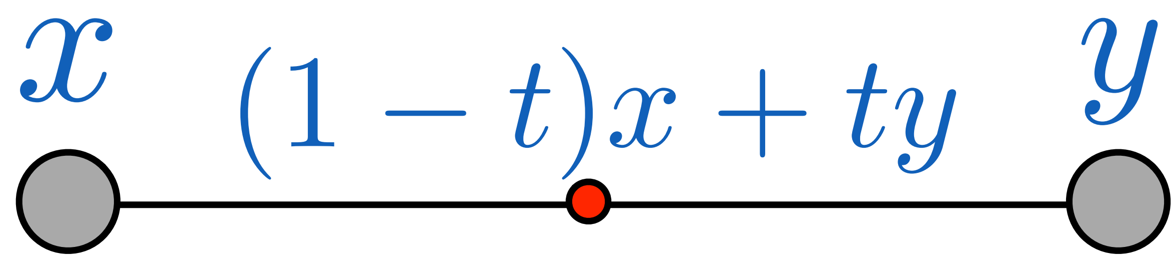


see also: [Rápcsák 1984; Udriste 1994]

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

The idea of geodesic convexity

Convexity



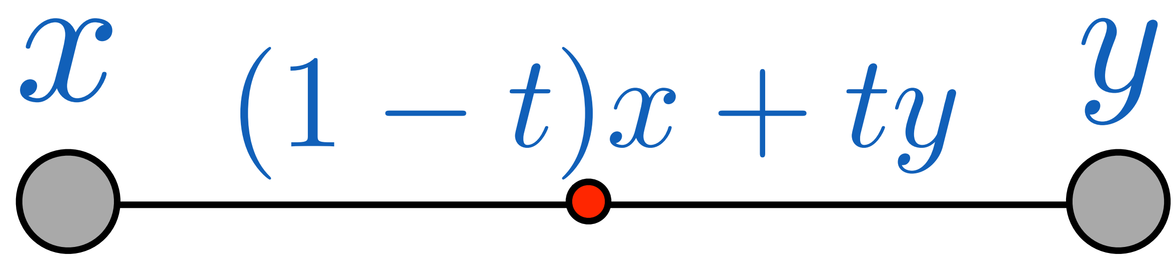
$$f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

see also: [Rápcsák 1984; Udriste 1994]

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

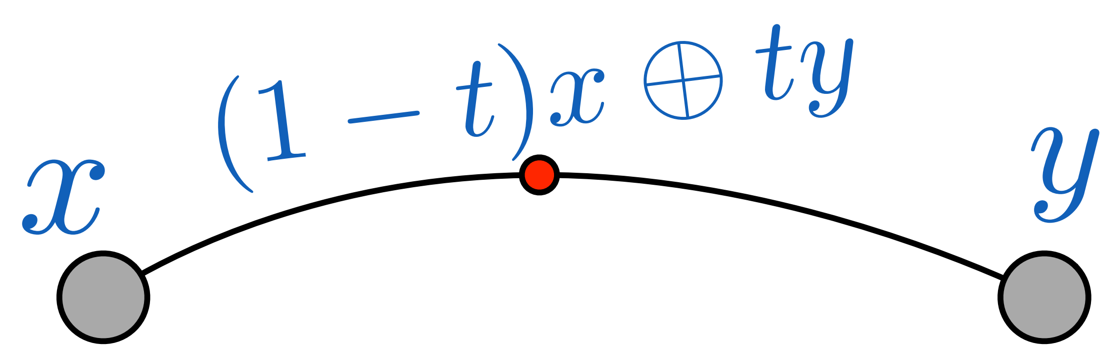
The idea of geodesic convexity

Convexity



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

Geodesic convexity

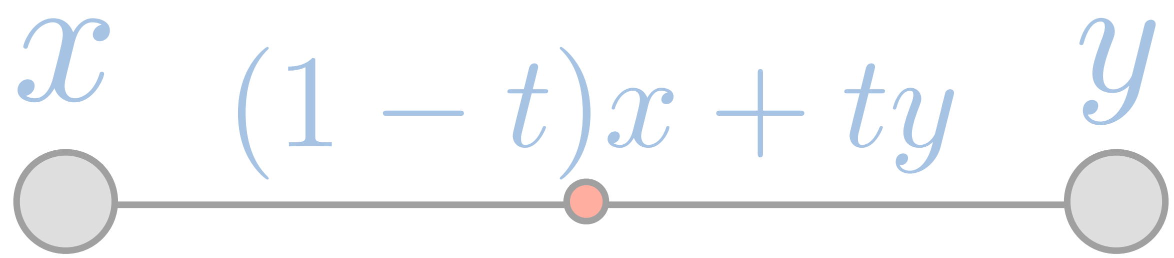


see also: [Rápcsák 1984; Udriste 1994]

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

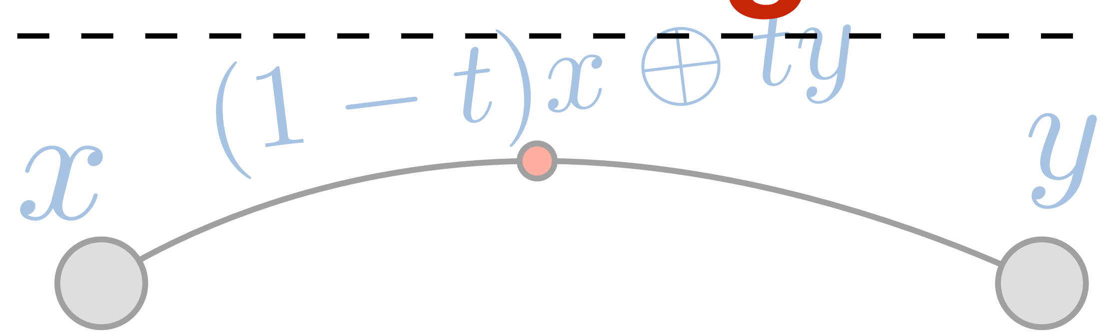
The idea of geodesic convexity

Convexity



Local opt of g-convex is global opt

Geodesic convexity

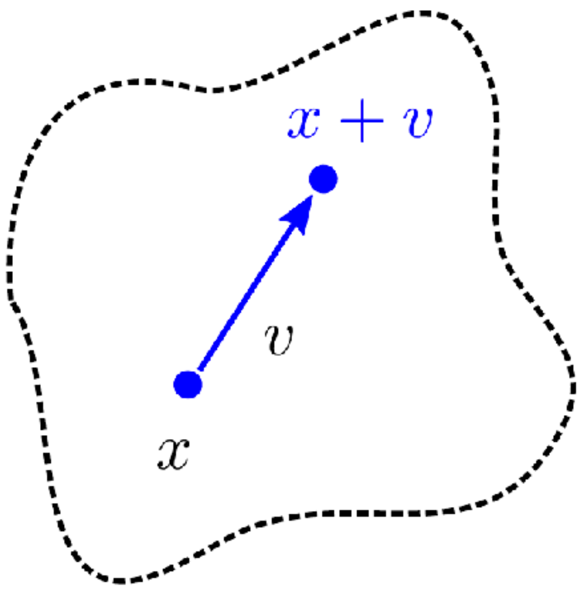


$$f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

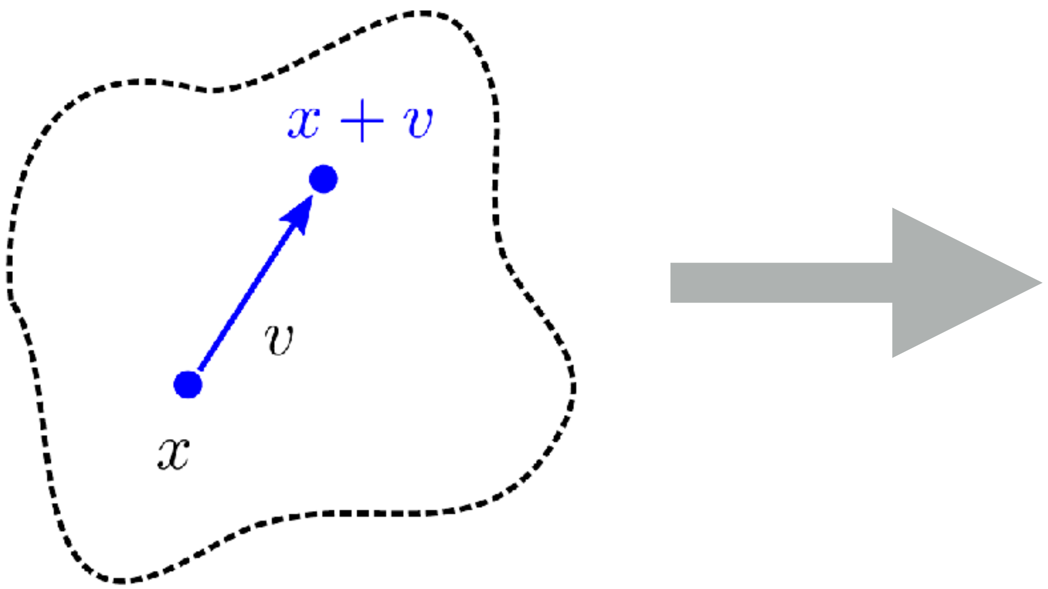
see also: [Rápcsák 1984; Udriste 1994]

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

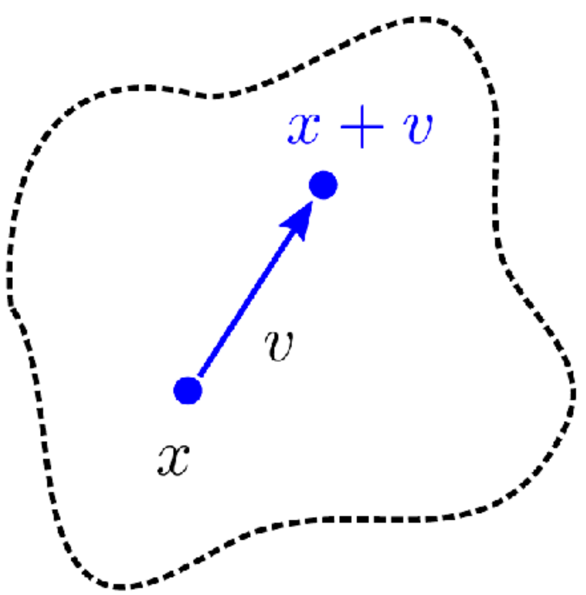
Key concepts generalize

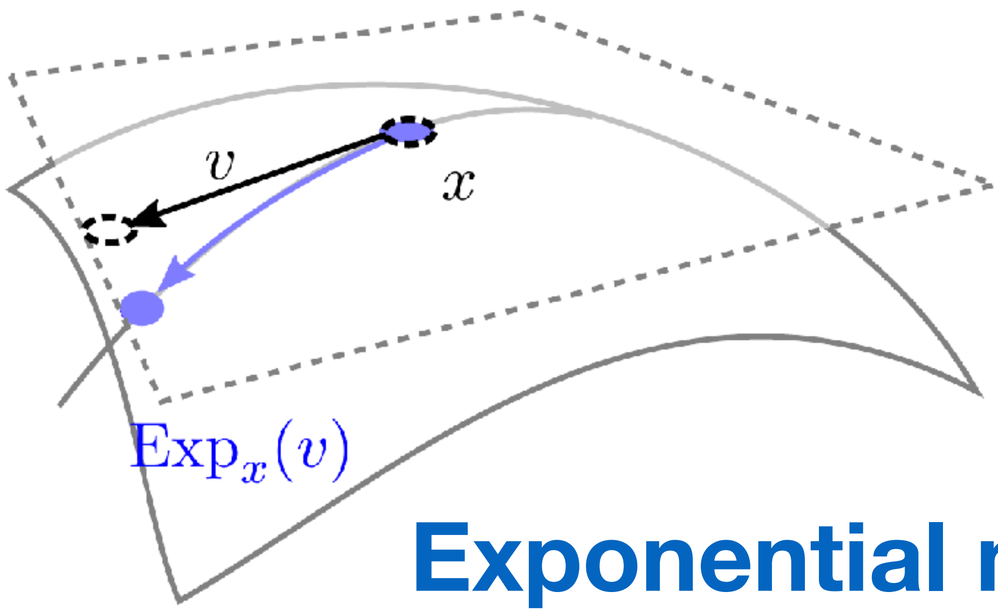


Key concepts generalize



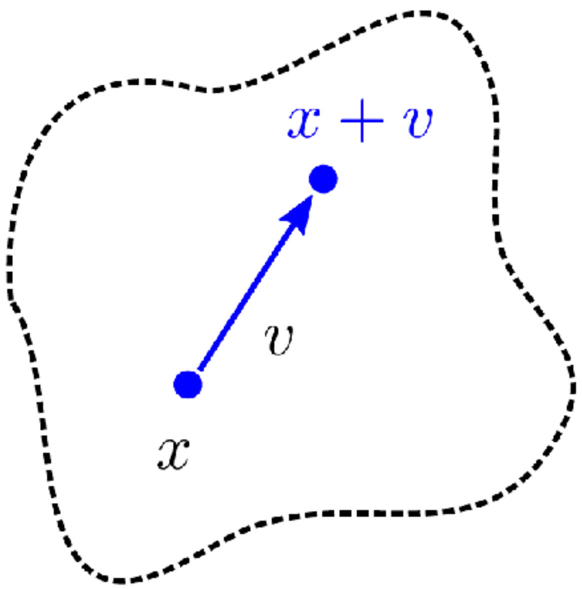
Key concepts generalize

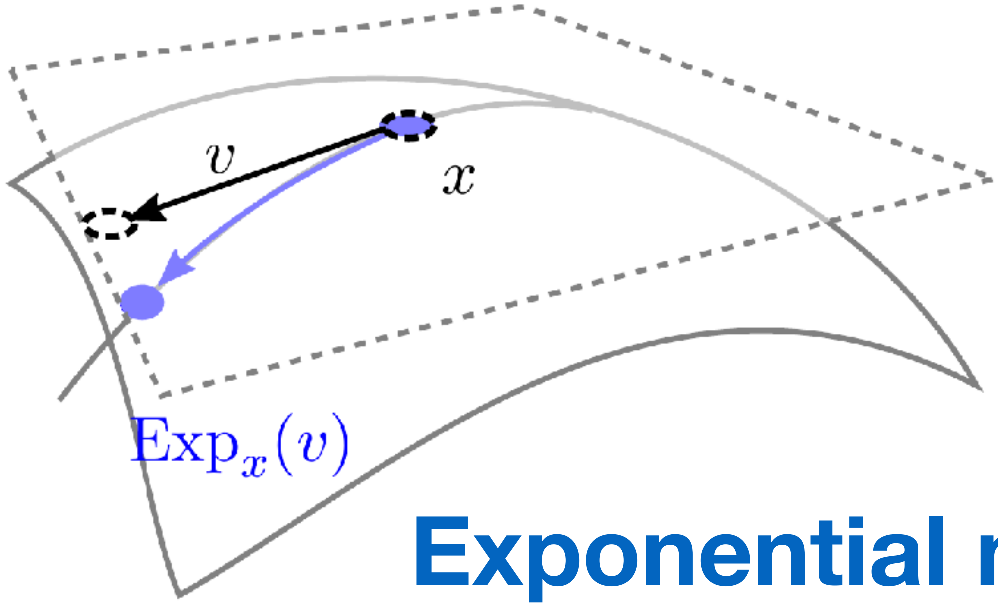




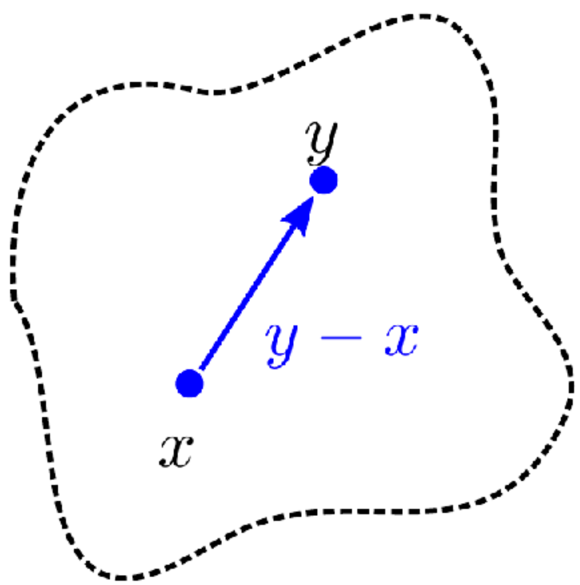
Exponential map

Key concepts generalize

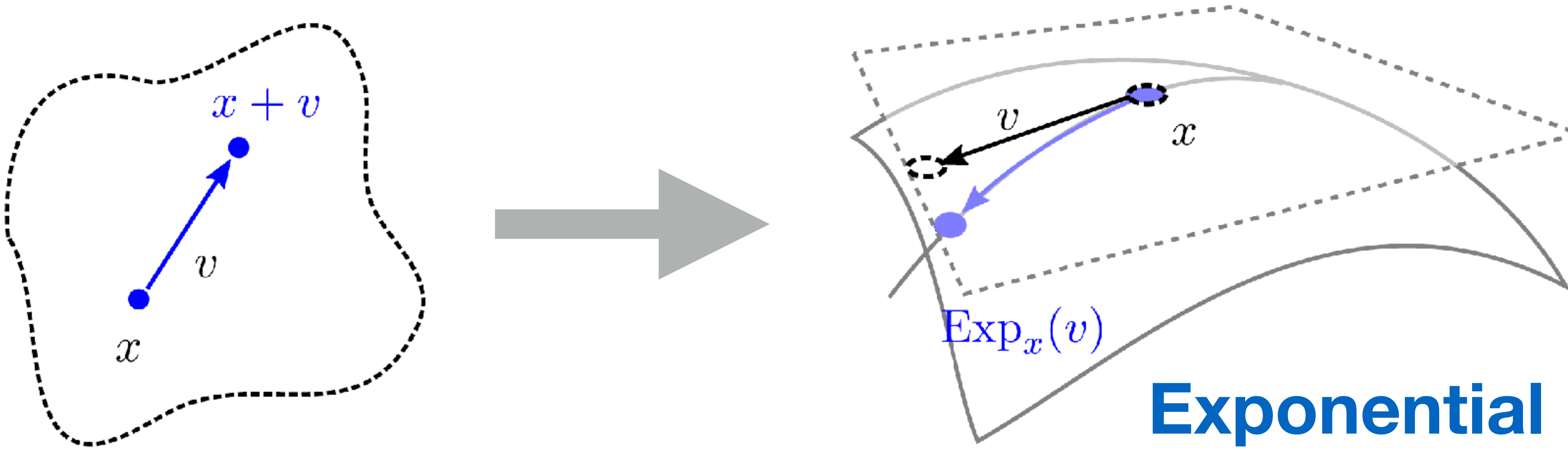




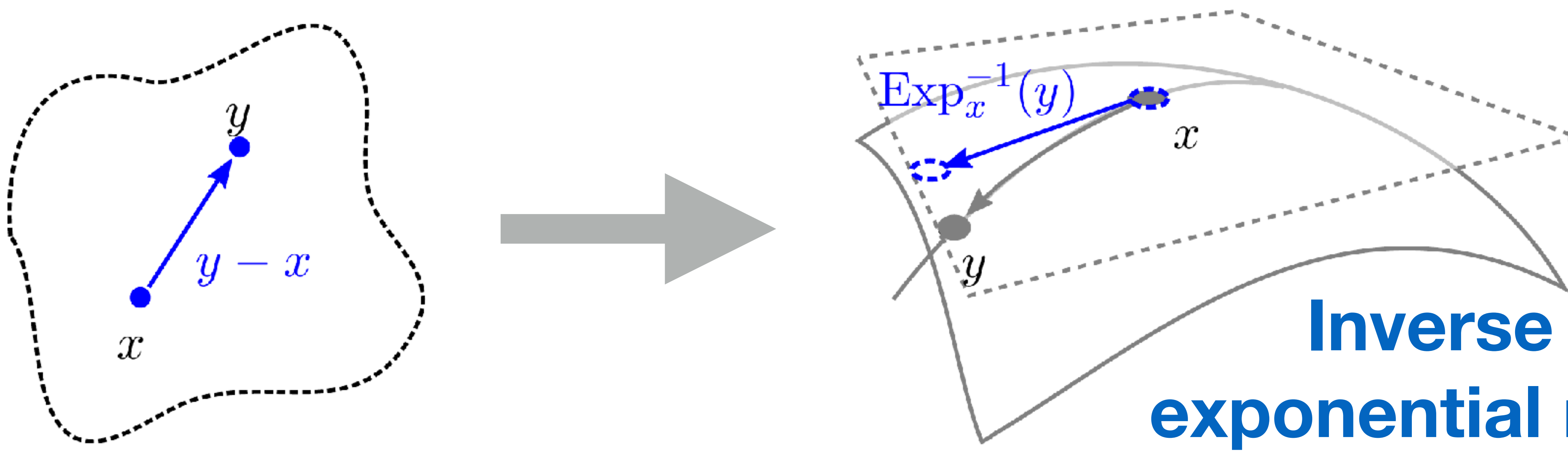
Exponential map



Key concepts generalize

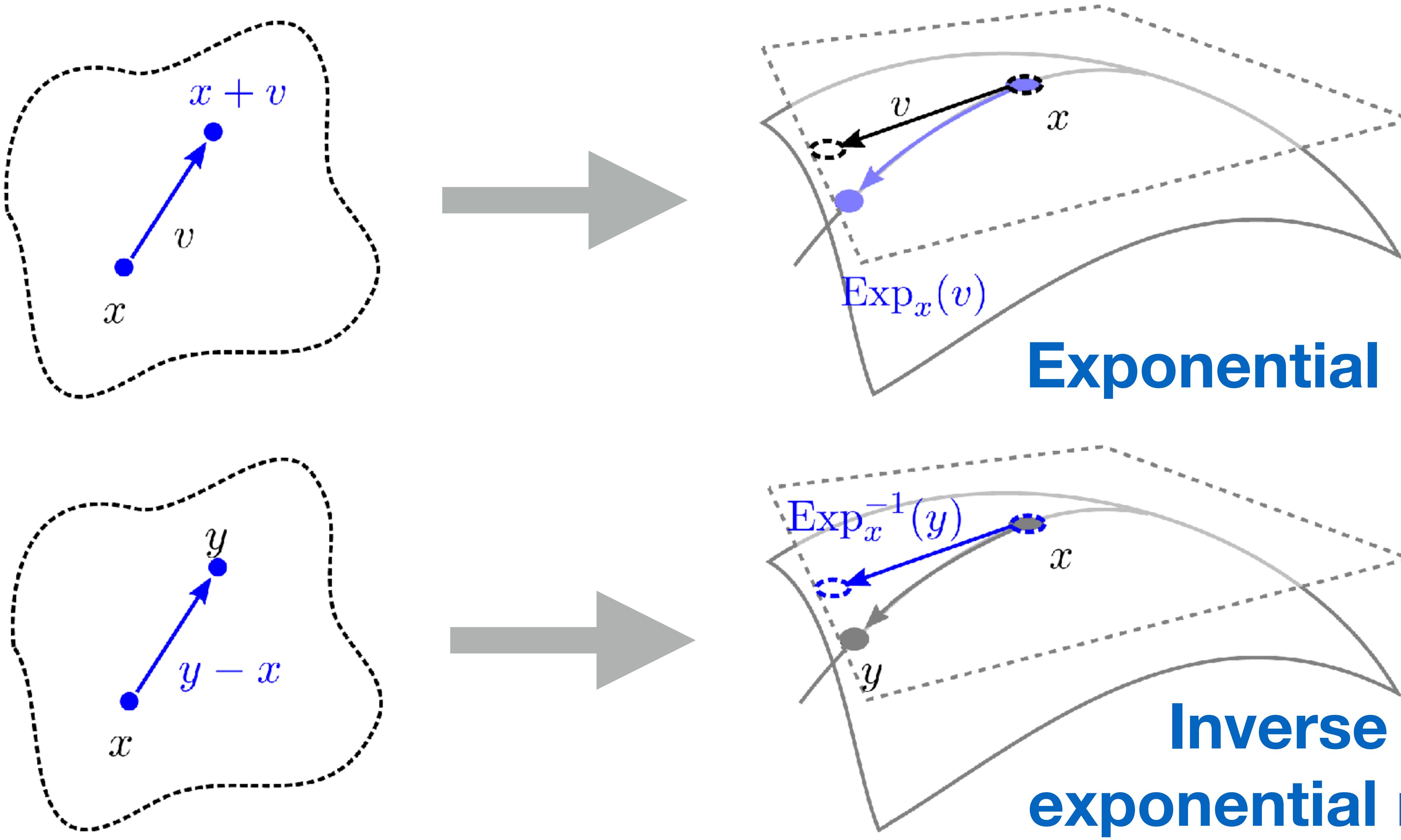


Exponential map



Inverse exponential map

Key concepts generalize



Exponential map

Inverse exponential map

lengths, angles, differentiation, vector translation, etc.

First-order algorithms

$$\min_{x \in \mathcal{X} \subset \mathcal{M}} \quad f(x)$$

First-order algorithms

$$\min_{x \in \mathcal{X} \subset \mathcal{M}} \quad f(x)$$

Assume: we can obtain exact or stochastic gradients

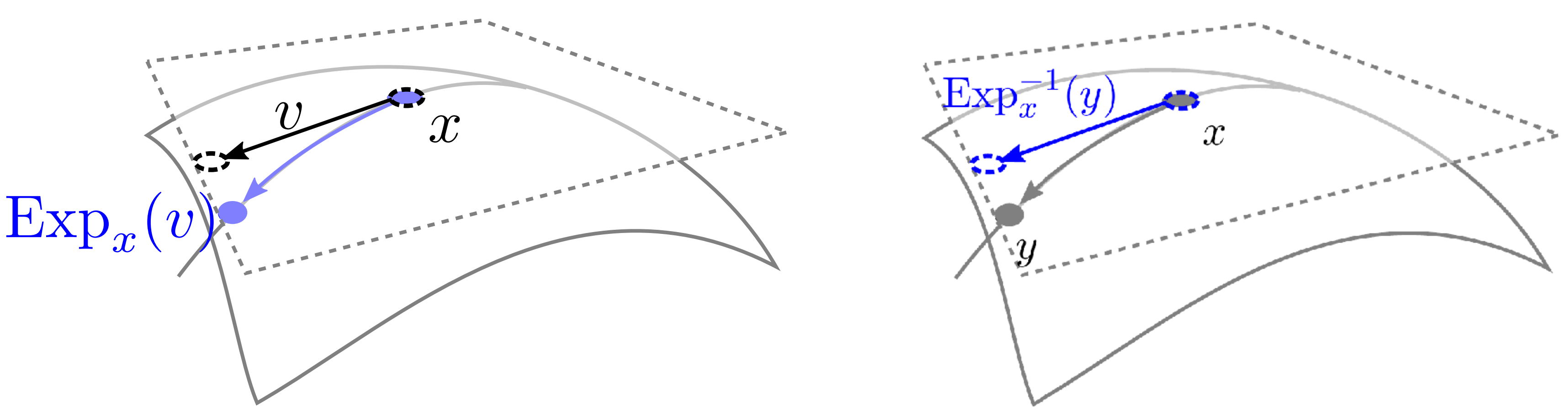
First-order algorithms

$$\min_{x \in \mathcal{X} \subset \mathcal{M}} \quad f(x)$$

Assume: we can obtain exact or stochastic gradients

Gradient descent $\quad x \leftarrow x - \eta \nabla f(x)$

GD on manifolds $\quad x \leftarrow \operatorname{Exp}_x(-\eta \nabla f(x))$



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 | Manifold Optimization

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

Example Manifolds

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}, \mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_\mathrm{F} = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$

[taken from manopt.org]

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}, \mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_{\mathrm{F}} = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_{\mathrm{F}} = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_{\mathrm{F}} = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$

[taken from manopt.org]

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_{\mathrm{F}} = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_{\mathrm{F}} = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_{\mathrm{F}} = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$

[taken from manopt.org]

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_\mathrm{F} = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$
Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^TX = I_p\}^k$
Complex Stiefel manifold	$\{X \in \mathbb{C}^{n\times p} : X^*X = I_p\}^k$
Generalized Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^TBX = I_p\}$ for some $B \succ 0$
Stiefel manifold, stacked	$\{X \in \mathbb{R}^{md\times k} : (XX^T)_{ii} = I_d\}$
Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^TX = I_p\}^k$
Complex Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{C}^{n\times p}, X^TX = I_p\}^k$
Generalized Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^TBX = I_p\}$ for some $B \succ 0$
Rotation group	$\{R \in \mathbb{R}^{n\times n} : R^TR = I_n, \det(R) = 1\}^k$
Special Euclidean group	$\{(R,t) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : R^TR = I_n, \det(R) = 1\}^k$

[taken from manopt.org]

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}$, $\mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_\mathrm{F} = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_\mathrm{F} = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$

Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^TX = I_p\}^k$
Complex Stiefel manifold	$\{X \in \mathbb{C}^{n\times p} : X^*X = I_p\}^k$
Generalized Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^TBX = I_p\}$ for some $B \succ 0$
Stiefel manifold, stacked	$\{X \in \mathbb{R}^{md\times k} : (XX^T)_{ii} = I_d\}$
Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^TX = I_p\}^k$
Complex Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{C}^{n\times p}, X^TX = I_p\}^k$
Generalized Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^TBX = I_p\}$ for some $B \succ 0$
Rotation group	$\{R \in \mathbb{R}^{n\times n} : R^TR = I_n, \det(R) = 1\}^k$
Special Euclidean group	$\{(R,t) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : R^TR = I_n, \det(R) = 1\}^k$

[taken from manopt.org]

Common Riemannian manifolds

Name	Set
Euclidean space (complex)	$\mathbb{R}^{m\times n}, \mathbb{C}^{m\times n}$
Symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T\}^k$
Skew-symmetric matrices	$\{X \in \mathbb{R}^{n\times n} : X + X^T = 0\}^k$
Centered matrices	$\{X \in \mathbb{R}^{m\times n} : X\mathbf{1}_n = 0_m\}$
Sphere	$\{X \in \mathbb{R}^{n\times m} : \|X\|_F = 1\}$
Symmetric sphere	$\{X \in \mathbb{R}^{n\times n} : \|X\|_F = 1, X = X^T\}$
Complex sphere	$\{X \in \mathbb{C}^{n\times m} : \|X\|_F = 1\}$
Oblique manifold	$\{X \in \mathbb{R}^{n\times m} : \|X_{:1}\| = \cdots = \|X_{:m}\| = 1\}$

Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^T X = I_p\}^k$
Complex Stiefel manifold	$\{X \in \mathbb{C}^{n\times p} : X^* X = I_p\}^k$
Generalized Stiefel manifold	$\{X \in \mathbb{R}^{n\times p} : X^T BX = I_p\}$ for some $B \succ 0$
Stiefel manifold, stacked	$\{X \in \mathbb{R}^{md\times k} : (XX^T)_{ii} = I_d\}$
Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^T X = I_p\}^k$
Complex Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{C}^{n\times p}, X^T X = I_p\}^k$
Generalized Grassmann manifold	$\{\mathrm{span}(X) : X \in \mathbb{R}^{n\times p}, X^T BX = I_p\}$ for some $B \succ 0$
Rotation group	$\{R \in \mathbb{R}^{n\times n} : R^T R = I_n, \det(R) = 1\}^k$
Special Euclidean group	$\{(R,t) \in \mathbb{R}^{n\times n} \times \mathbb{R}^n : R^T R = I_n, \det(R) = 1\}^k$

[taken from manopt.org]

Common Riemannian manifolds

Fixed-rank	$\{X \in \mathbb{R}^{m \times n} : \operatorname{rank}(X) = k\}$
Fixed-rank tensor	Tensors of fixed multilinear rank in Tucker format
Matrices with strictly positive entries	$\{X \in \mathbb{R}^{m \times n} : X_{ij} > 0 \ \forall i, j\}$
Symmetric, positive definite matrices	$\{X \in \mathbb{R}^{n \times n} : X = X^T, X \succ 0\}^k$
Symmetric positive semidefinite, fixed-rank	$\{X \in \mathbb{R}^{n \times n} : X = X^T \succeq 0, \operatorname{rank}(X) = k\}$

[taken from manopt.org]

Common Riemannian manifolds

Fixed-rank	$\{X \in \mathbb{R}^{m\times n} : \text{rank}(X) = k\}$
Fixed-rank tensor	Tensors of fixed multilinear rank in Tucker format
Matrices with strictly positive entries	$\{X \in \mathbb{R}^{m\times n} : X_{ij} > 0\ \forall i,j\}$
Symmetric, positive definite matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T, X \succ 0\}^k$
Symmetric positive semidefinite, fixed-rank	$\{X \in \mathbb{R}^{n\times n} : X = X^T \succeq 0, \text{rank}(X) = k\}$

[taken from manopt.org]

Common Riemannian manifolds

Fixed-rank	$\{X \in \mathbb{R}^{m \times n} : \operatorname{rank}(X) = k\}$
Fixed-rank tensor	Tensors of fixed multilinear rank in Tucker format
Matrices with strictly positive entries	$\{X \in \mathbb{R}^{m \times n} : X_{ij} > 0 \ \forall i, j\}$
Symmetric, positive definite matrices	$\{X \in \mathbb{R}^{n \times n} : X = X^T, X \succ 0\}^k$
Symmetric positive semidefinite, fixed-rank	$\{X \in \mathbb{R}^{n \times n} : X = X^T \succeq 0, \operatorname{rank}(X) = k\}$

[taken from manopt.org]

Common Riemannian manifolds

Fixed-rank	$\{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = k\}$
Fixed-rank tensor	Tensors of fixed multilinear rank in Tucker fo
Matrices with strictly positive entries	$\{X \in \mathbb{R}^{m\times n} : X_{ij} > 0\ \forall i, j\}$
Symmetric, positive definite matrices	$\{X \in \mathbb{R}^{n\times n} : X = X^T, X \succ 0\}^k$
Symmetric positive semidefinite, fixed-rank	$\{X \in \mathbb{R}^{n\times n} : X = X^T \succeq 0, \mathrm{rank}(X) = k\}$
Symmetric positive semidefinite, fixed-rank with unit diagonal	$\{X \in \mathbb{R}^{n\times n} : X = X^T \succeq 0, \mathrm{rank}(X) = k, \mathrm{diag}(X) = 1\}$
Symmetric positive semidefinite, fixed-rank with unit trace	$\{X \in \mathbb{R}^{n\times n} : X = X^T \succeq 0, \mathrm{rank}(X) = k, \mathrm{trace}(X) = 1\}$
Multinomial manifold (strict simplex elements)	$\{X \in \mathbb{R}^{n\times m} : X_{ij} > 0 \forall i, j \text{ and } X^T \mathbf{1}_m = \mathbf{1}_n\}$
Multinomial doubly stochastic manifold	$\{X \in \mathbb{R}^{n\times n} : X_{ij} > 0 \forall i, j \text{ and } X\mathbf{1}_n = \mathbf{1}_n, X^T\mathbf{1}_n = \mathbf{1}_n\}$
Multinomial symmetric and stochastic manifold	$\{X \in \mathbb{R}^{n\times n} : X_{ij} > 0 \forall i, j \text{ and } X\mathbf{1}_n = \mathbf{1}_n, X = X^T\}$

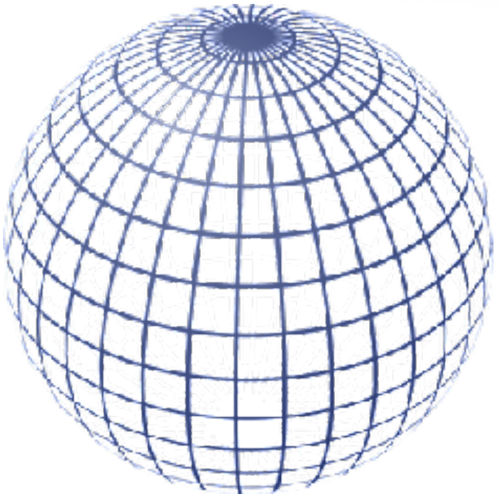
[taken from manopt.org]

Examples & Applications

Eigenvector problems

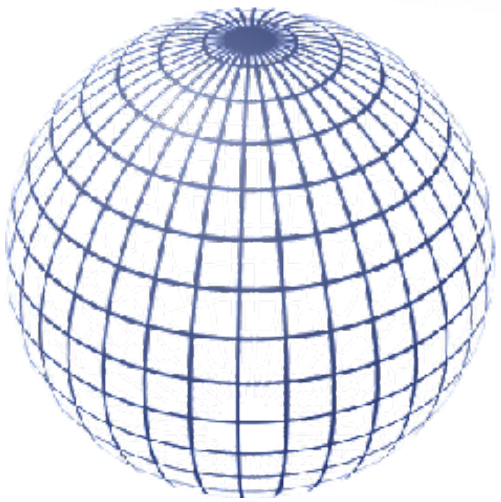
Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Eigenvector problems



Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Eigenvector problems

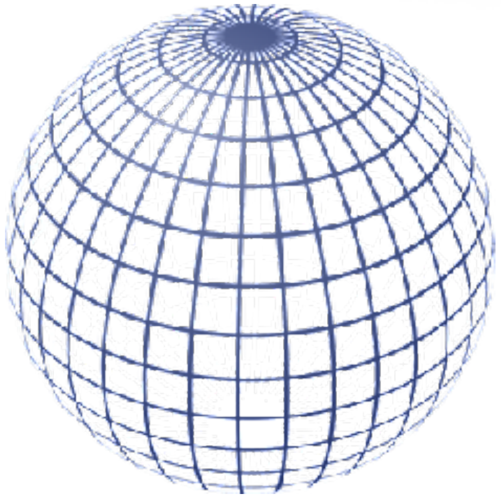


Largest eigenvector

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

Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Eigenvector problems



Largest eigenvector

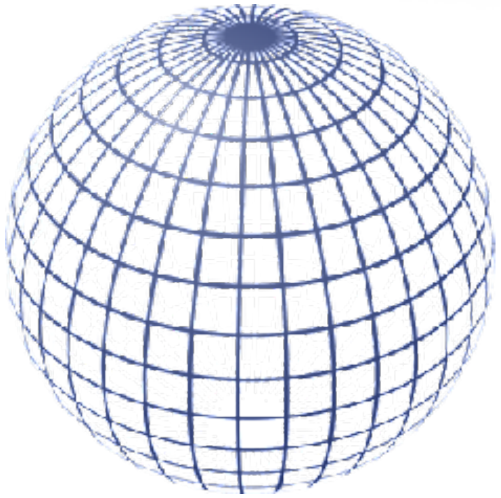
$$\max_{x^T x=1} x^T A x$$

Power iteration

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

Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Eigenvector problems



Largest eigenvector

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

Power iteration

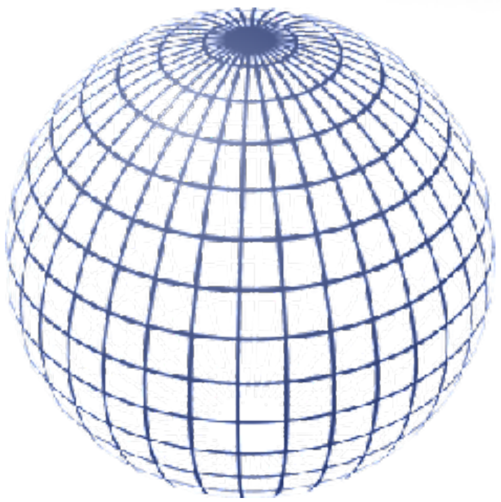
$$\boxed{x \leftarrow \frac{Ax}{\|Ax\|}}$$

May be viewed as Riemannian gradient descent

(albeit under another "retraction" instead of the Exp-map)

Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Eigenvector problems



Largest eigenvector

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

Power iteration

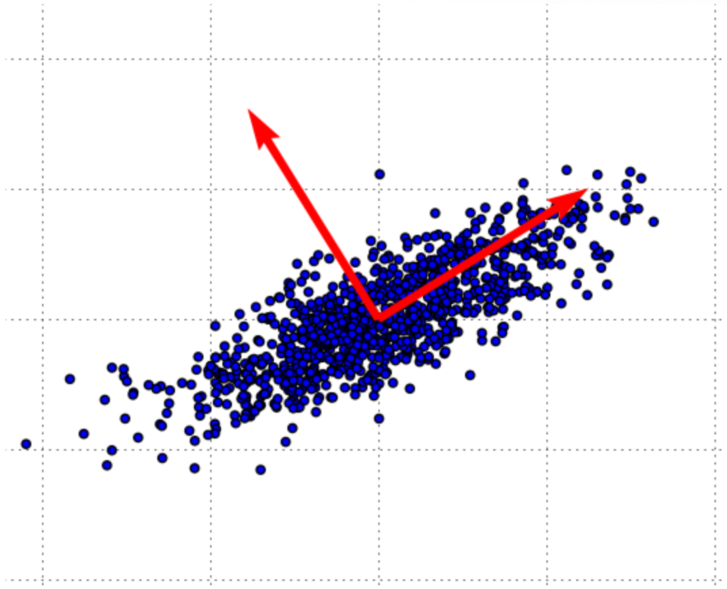
$$\boxed{x \leftarrow \frac{Ax}{\|Ax\|}}$$

May be viewed as Riemannian gradient descent
(albeit under another "retraction" instead of the Exp-map)

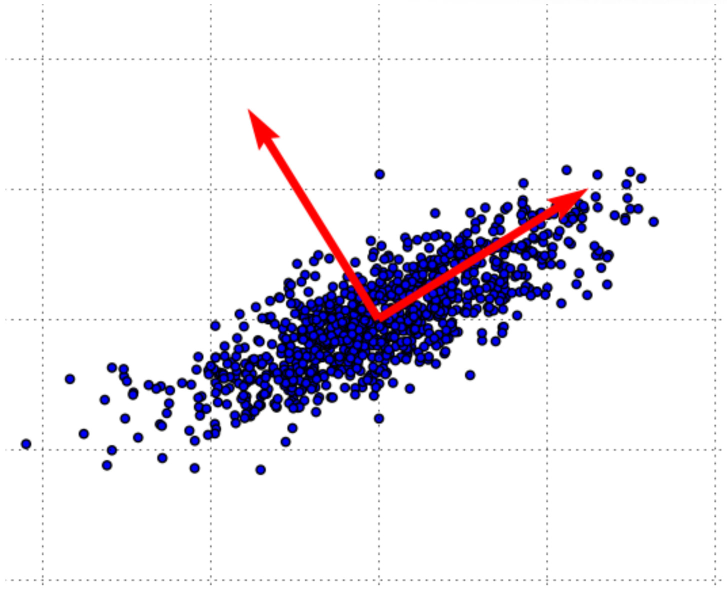
Provides transparent reasoning for global convergence rate of iter

Ref: *Projection-like retractions on matrix manifolds,* Pierre-Antoine Absil, Jérôme Malick.

Stochastic eigenvectors (large-scale)



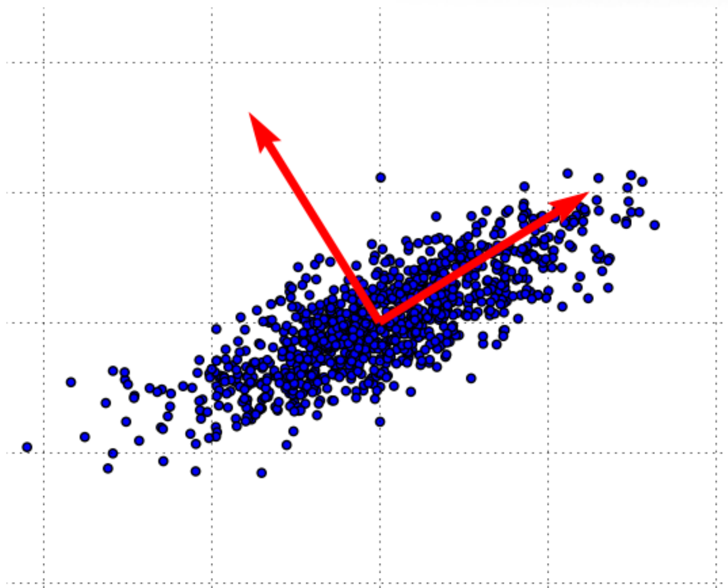
Stochastic eigenvectors (large-scale)



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

n is big

Stochastic eigenvectors (large-scale)



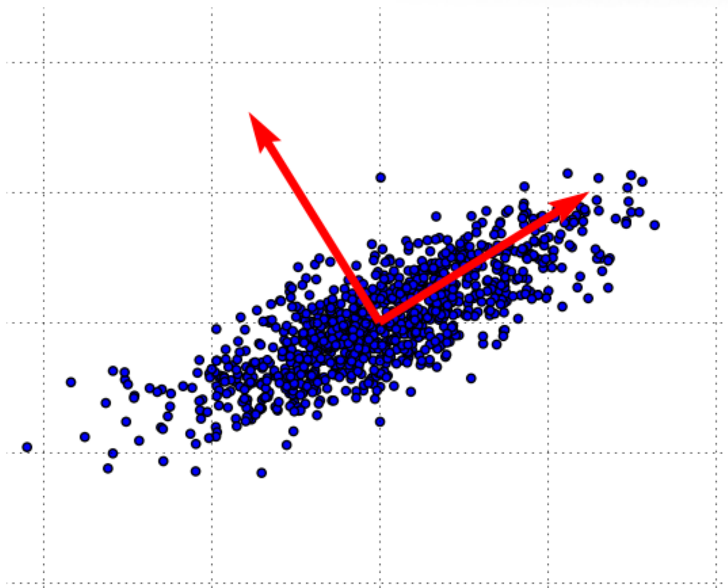
$$\min_{x^T x = 1} \quad -x^T \left(\sum_{i=1}^{n} z_i z_i^T\right) x$$

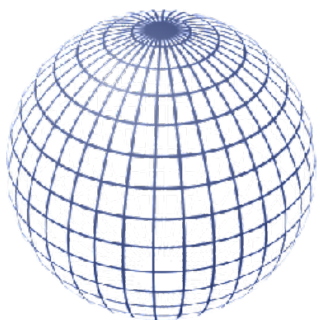
n is big

Lots of work on "SGD" for eigenvectors exists

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

Stochastic eigenvectors (large-scale)





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

n is big

Lots of work on "SGD" for eigenvectors exists

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

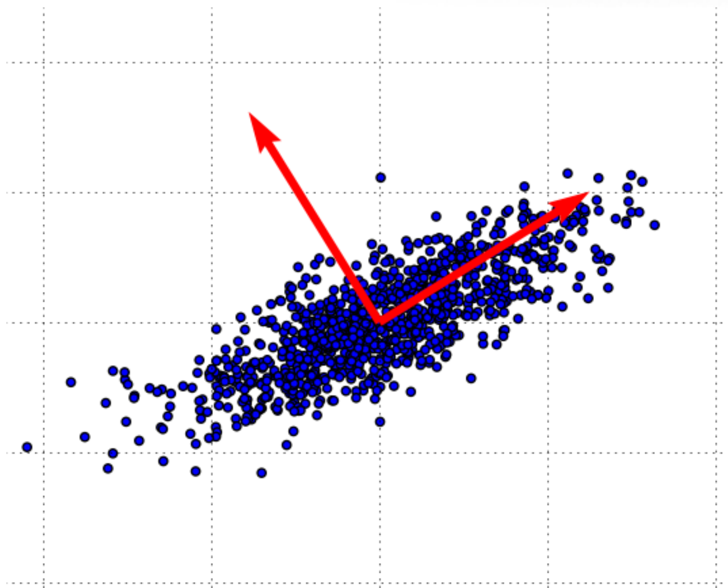
Simpler analysis thanks to a key geometric realization

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

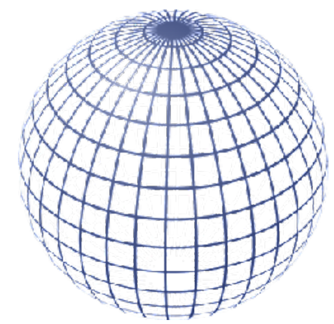
Running Riemannian SGD will obtain global optimum

[Zhang, Reddi, Sra, NIPS 2016]

Stochastic eigenvectors (large-scale)



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



n is big

Lots of work on “SGD” for eigenvectors exists

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

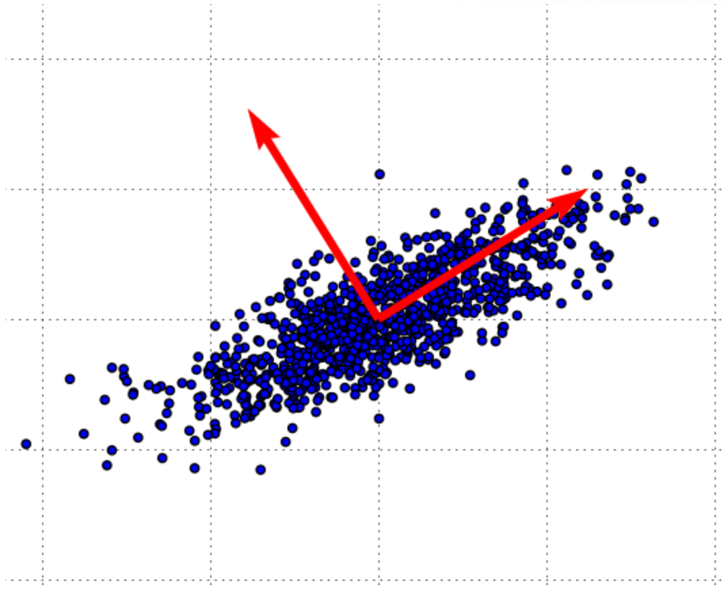
Simpler analysis thanks to a key geometric realization

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 $O(\frac{d}{p^2\delta})$-gradient dominated.

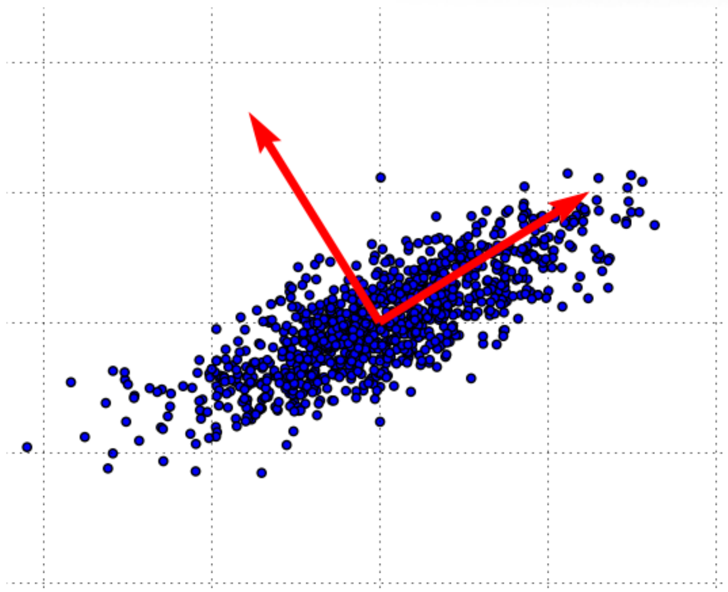
Running Riemannian SGD will obtain global optimum

[Zhang, Reddi, Sra, NIPS 2016]

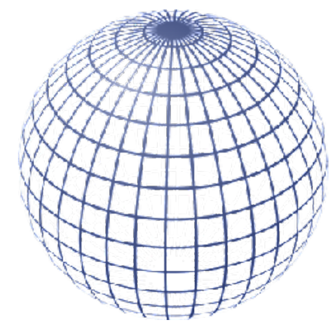
Stochastic eigenvectors (large-scale)



Stochastic eigenvectors (large-scale)

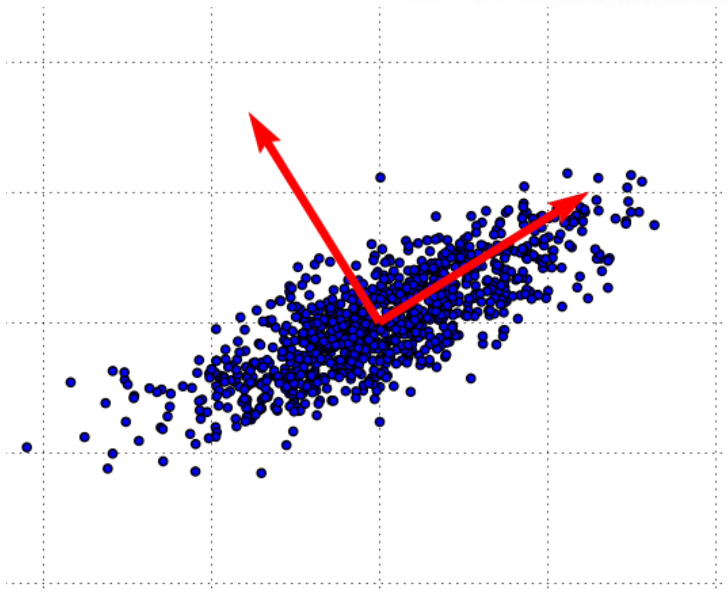


$$\min_{x^Tx=1} \quad -x^T\left(\sum_{i=1}^{n} z_i z_i^T\right)x$$

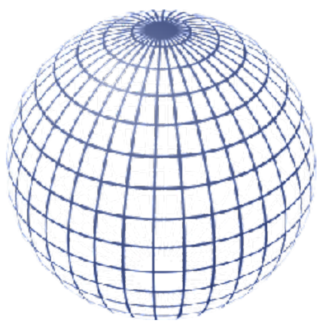


n is big

Stochastic eigenvectors (large-scale)



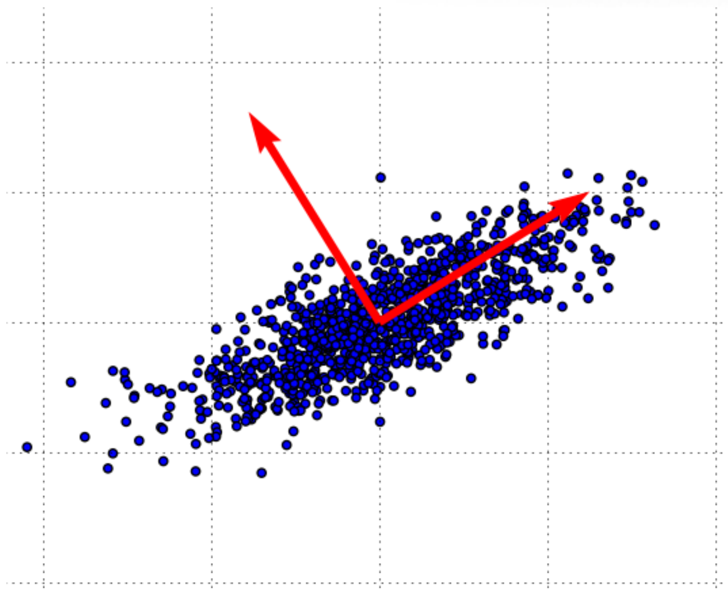
$$\min_{x^Tx=1} \; -x^T\left(\sum_{i=1}^n z_i z_i^T\right)x$$



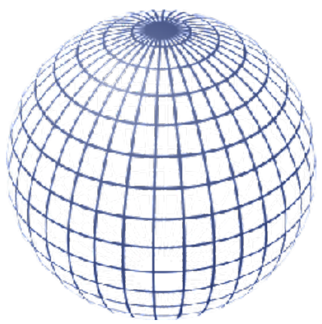
n is big

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 $O(\frac{d}{p^2\delta})$-gradient dominated.

Stochastic eigenvectors (large-scale)



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

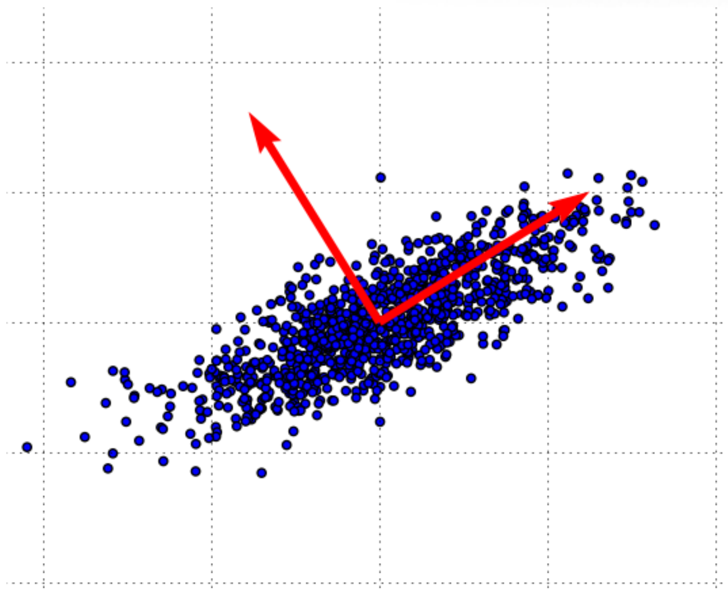


n is big

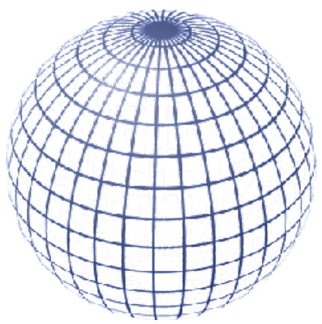
Theorem 4. Suppose A has eigenvalues $\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_d$ and $\delta = \lambda_1 - \lambda_2$. With probability $1-p$, the random initialization x^0 falls in a Riemannian ball of a global optimum of the objective function, within which the objective function is $O(\frac{d}{p^2\delta})$-gradient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

1. A more careful initialization should improve the bound

Stochastic eigenvectors (large-scale)



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



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 $O\left(\frac{d}{p^2 \delta}\right)$-gradient dominated.

1. A more careful initialization should improve the bound

2. Can we accelerate to $\sqrt{\delta}$?

Matrix Factorization

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

Matrix Factorization

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

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

Matrix Factorization

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

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

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

Nicolas Boumal[a,*], P.-A. Absil[b]

Matrix Factorization

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

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

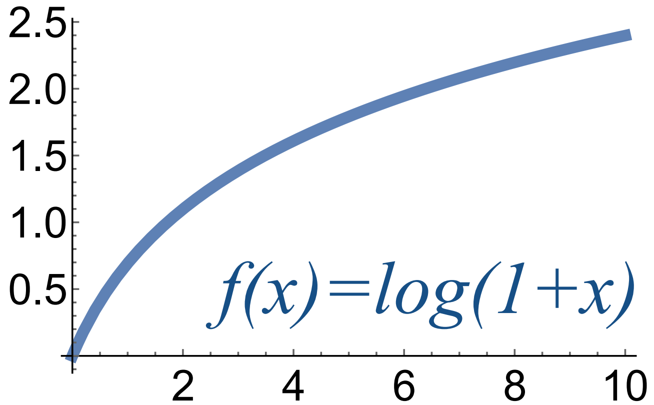
Low-rank matrix completion via preconditioned optimization on the Grassmann manifold

Nicolas Boumal [a,*], P.-A. Absil [b]

Riemannian Perspective on Matrix Factorization

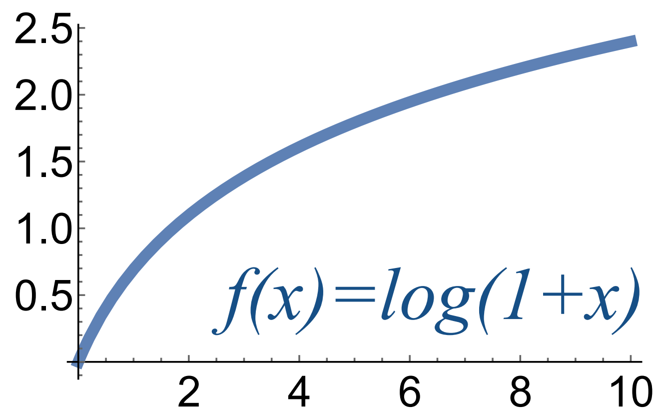
Kwangjun Ahn [*1] and Felipe Suarez [†2]

G-convexity for positive definite matrices



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

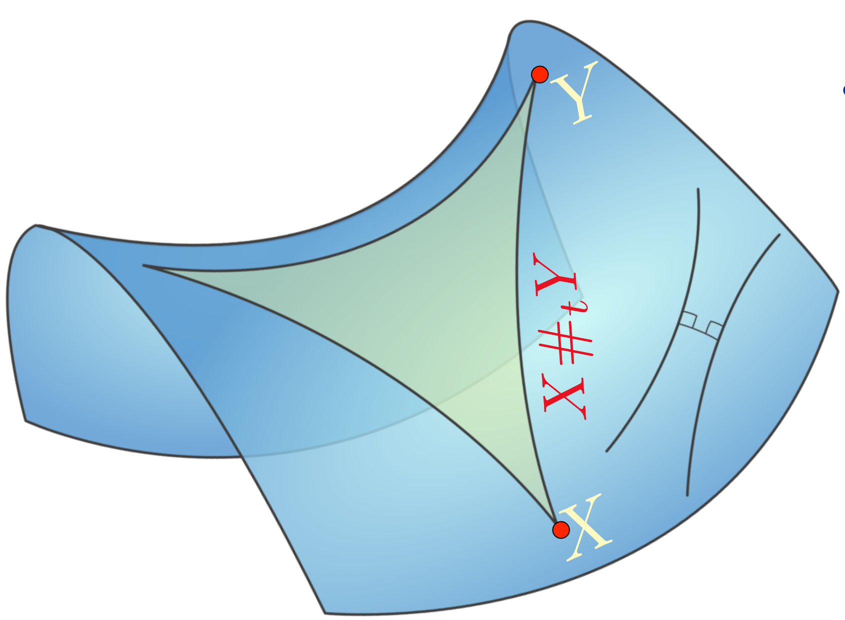
G-convexity for positive definite matrices



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

$$\gamma(t) \equiv (1-t)X \oplus tY := X^{\frac{1}{2}}(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})^t X^{\frac{1}{2}}$$



$$f((1-t)X \oplus tY) \leqslant (1-t)f(X) + tf(Y)$$

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, Nemirovksi 1991]

log-g-convex

Trace of power

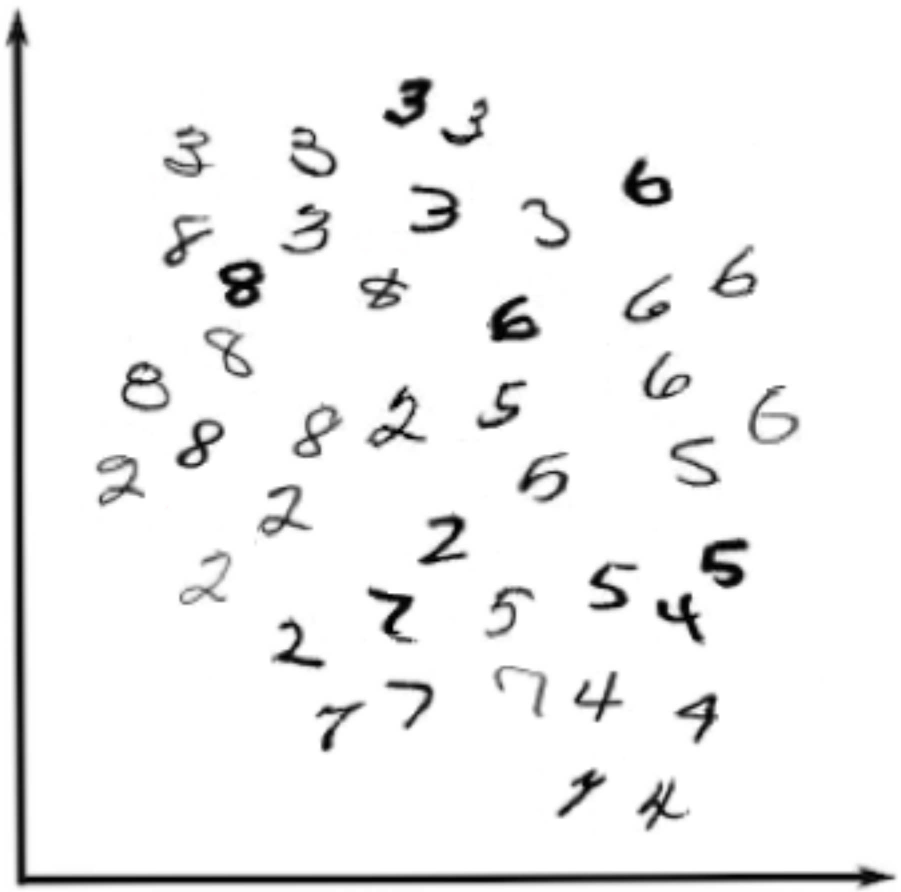
$$\log \operatorname{tr}(X^p), \quad p \in \mathbb{R}$$

and many more…

[Sra 2017]

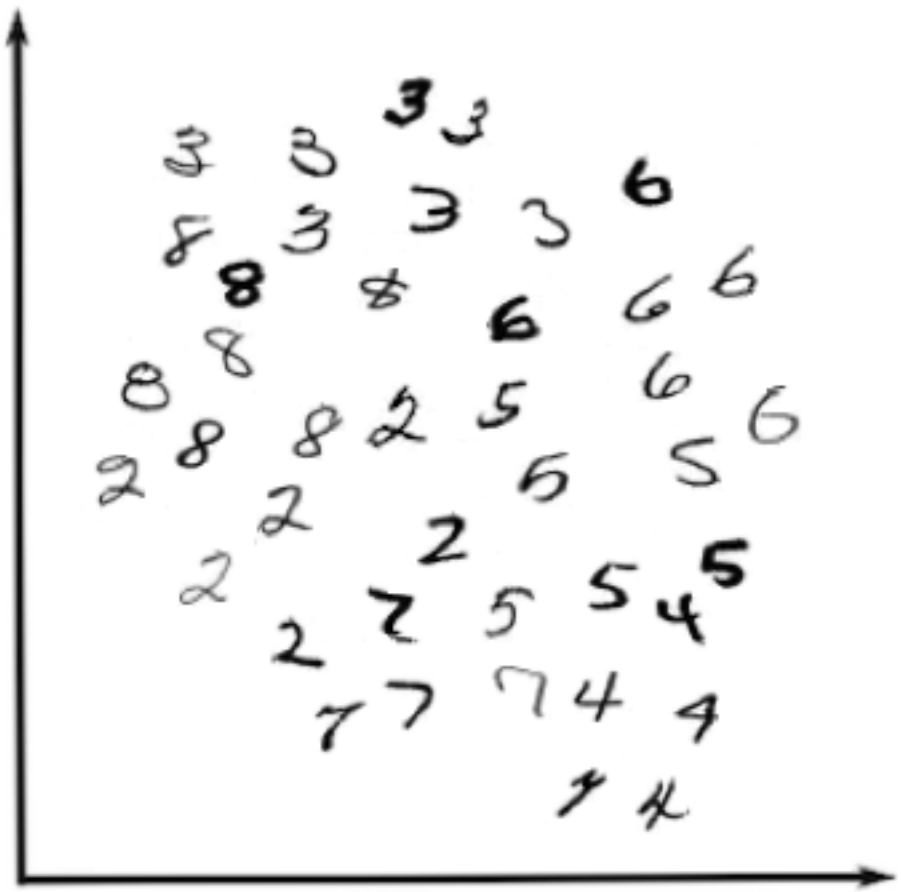
Example: metric learning

Metric learning: a fundamental problem in machine learning



Example: metric learning

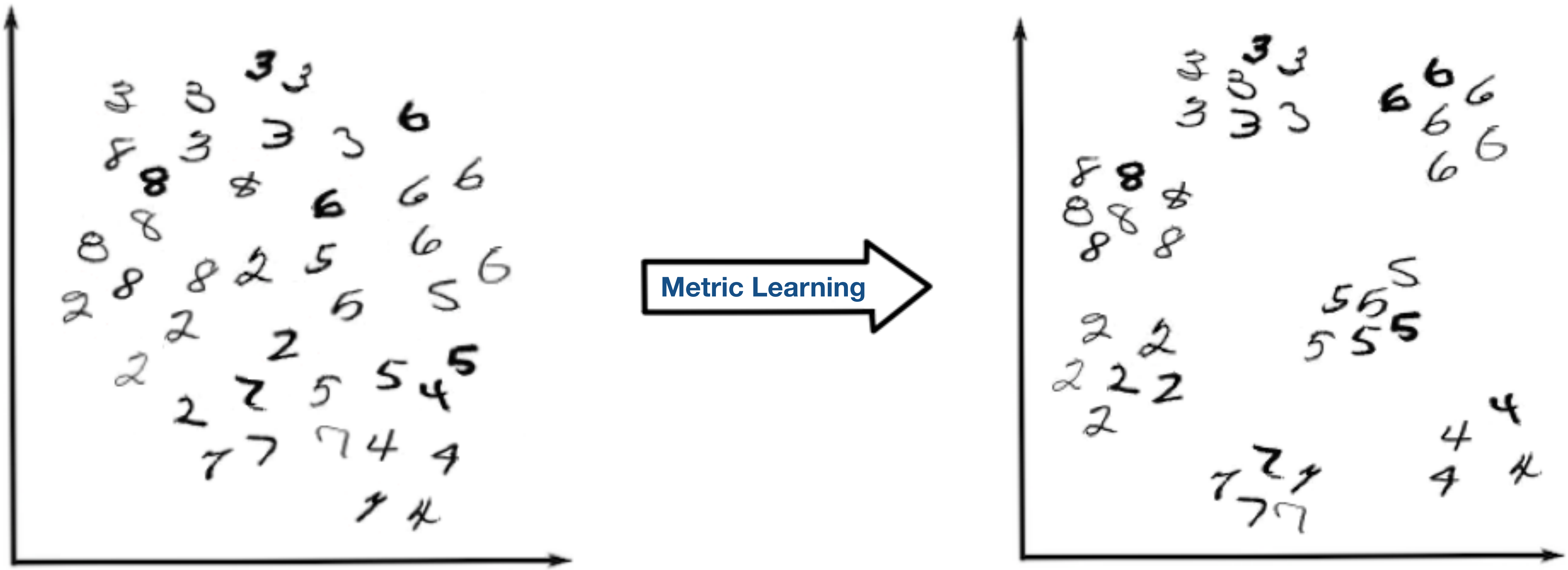
Metric learning: a fundamental problem in machine learning



If we can judge "similarity" between data points, classification becomes easy (eg via nearest neighbors)

Example: metric learning

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*

$$\mathcal{S} := \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in the same class}\}$$
$$\mathcal{D} := \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in different classes}\}$$

Goal: *learn Mahalanobis distance*

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})$$

Ensure*: distances between similar points are small*
distances between dissimilar points are large

Linear metric learning

Input: *pairwise constraints*

$$\mathcal{S} := \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in the same class}\}$$
$$\mathcal{D} := \{(\boldsymbol{x}_i, \boldsymbol{x}_j) \mid \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are in different classes}\}$$

Goal: *learn Mahalanobis distance*

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boxed{\boldsymbol{A}} (\boldsymbol{x} - \boldsymbol{y})$$

Ensure*: distances between similar points are small*
distances between dissimilar points are large

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A} (\boldsymbol{x} - \boldsymbol{y})$$

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

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

$$\text{such that} \quad \sum_{(x_i, x_j) \in \mathcal{D}} \sqrt{d_A(x_i, x_j)} \geq 1$$

$$d_A(x, y) := (x - y)^T A (x - y)$$

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

$$\min_{\boldsymbol{A}\succeq 0} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

$$\text{such that} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}} \sqrt{d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)} \geq 1$$

LMNN

[Weinberger, Saul 2005]

large-margin SDP

$$\min_{\boldsymbol{A}\succeq 0} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} \Big[(1-\mu)d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \mu \sum\nolimits_l (1-y_{il})\xi_{ijl}\Big]$$

$$d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_l) - d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j) \geq 1-\xi_{ijl}$$

$$\xi_{ijl} \geq 0$$

$$d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y}) := (\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})$$

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

$$\min_{\boldsymbol{A}\succeq 0} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$
$$\text{such that} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{D}} \sqrt{d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j)} \geq 1$$

LMNN

[Weinberger, Saul 2005]

large-margin SDP

$$\min_{\boldsymbol{A}\succeq 0} \sum_{(\boldsymbol{x}_i,\boldsymbol{x}_j)\in\mathcal{S}} \Big[(1-\mu)d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j) + \mu \textstyle\sum_l (1-y_{il})\xi_{ijl}\Big]$$
$$d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_l) - d_{\boldsymbol{A}}(\boldsymbol{x}_i,\boldsymbol{x}_j) \geq 1 - \xi_{ijl}$$
$$\xi_{ijl} \geq 0$$

ITML

[Davis, Kulis, Jain, Sra, Dhillon 2007]

relative entropy b/w Gaussians

$$\begin{aligned} \min_{\boldsymbol{A}\succeq 0} \quad & D_{\mathrm{ld}}(\boldsymbol{A},\boldsymbol{A}_0) \\ \text{such that} \quad & d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y}) \leq u, \quad (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{S}, \\ & d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y}) \geq l, \quad (\boldsymbol{x},\boldsymbol{y}) \in \mathcal{D} \end{aligned}$$

$$D_{\mathrm{ld}}(\boldsymbol{A},\boldsymbol{A}_0) := \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - \log\det(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - d$$

$$d_{\boldsymbol{A}}(\boldsymbol{x},\boldsymbol{y}) := (\boldsymbol{x}-\boldsymbol{y})^T\boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})$$

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

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

$$\text{such that} \quad \sum_{(x_i, x_j) \in \mathcal{D}} \sqrt{d_A(x_i, x_j)} \geq 1$$

LMNN

[Weinberger, Saul 2005]

large-margin SDP

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

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

$$\xi_{ijl} \geq 0$$

ITML

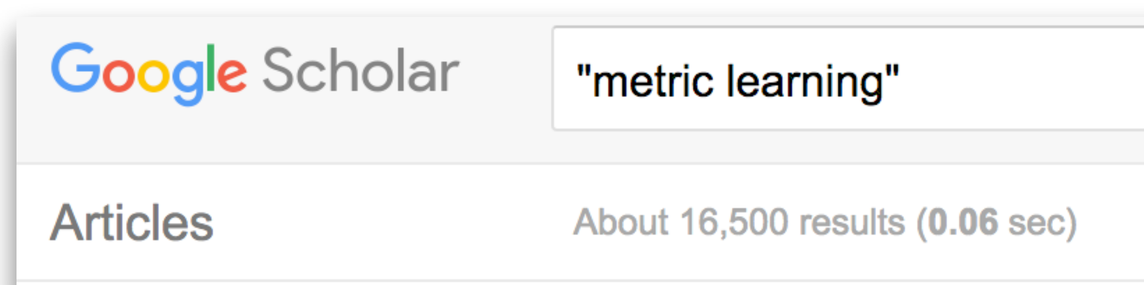
[Davis, Kulis, Jain, Sra, Dhillon 2007]

relative entropy b/w Gaussians

$$\begin{aligned} \min_{A \succeq 0} \quad & D_{\text{ld}}(A, A_0) \\ \text{such that} \quad & d_A(x, y) \leq u, \quad (x, y) \in \mathcal{S}, \\ & d_A(x, y) \geq l, \quad (x, y) \in \mathcal{D} \end{aligned}$$

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

Tons of other works



$$d_A(x, y) := (x - y)^T A (x - y)$$

A new geometric approach

$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})$

Euclidean idea

$$\min_{\boldsymbol{A} \succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

A new geometric approach

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})$$

Euclidean idea

$$\min_{\boldsymbol{A} \succeq 0} \cancel{\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}$$

A new geometric approach

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x} - \boldsymbol{y})^T \boldsymbol{A}(\boldsymbol{x} - \boldsymbol{y})$$

Euclidean idea

$$\cancel{\min_{\boldsymbol{A} \succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}$$

New idea

$$\min_{\boldsymbol{A} \succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}^{-1}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

A new geometric approach

$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) := (\boldsymbol{x}-\boldsymbol{y})^T \boldsymbol{A}(\boldsymbol{x}-\boldsymbol{y})$

Euclidean idea

$$\min_{\boldsymbol{A} \succeq 0} \cancel{\sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)}$$

New idea

$$\min_{\boldsymbol{A} \succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}^{-1}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Intuitively: If $a > b$, then $a^{-1} < b^{-1}$

Geometric approach to metric learning

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

$$\boldsymbol{S} := \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T,$$

$$\boldsymbol{D} := \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$$

scatter matrices

[Habibzadeh, Hosseini, Sra, ICML 2016]

Geometric approach to metric learning

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

$$\boldsymbol{S} := \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T,$$

$$\boldsymbol{D} := \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} (\boldsymbol{x}_i - \boldsymbol{x}_j)(\boldsymbol{x}_i - \boldsymbol{x}_j)^T$$

scatter matrices

Equivalently solve

$$\min_{\boldsymbol{A} \succ 0} \quad h(\boldsymbol{A}) := \operatorname{tr}(\boldsymbol{A}\boldsymbol{S}) + \operatorname{tr}(\boldsymbol{A}^{-1}\boldsymbol{D})$$

[Habibzadeh, Hosseini, Sra, ICML 2016]

Geometric approach to metric learning

Closed form solution!

$$\nabla h(\boldsymbol{A}) = 0 \quad \Leftrightarrow \quad \boldsymbol{S} - \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} = 0$$

Geometric approach to metric learning

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

Closed form solution!

$$\nabla h(\boldsymbol{A}) = 0 \quad \Leftrightarrow \quad \boldsymbol{S} - \boldsymbol{A}^{-1} \boldsymbol{D} \boldsymbol{A}^{-1} = 0$$

$$\boldsymbol{A} = \boldsymbol{S}^{-1} \#_{\frac{1}{2}} \boldsymbol{D}$$

Geometric approach to metric learning

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

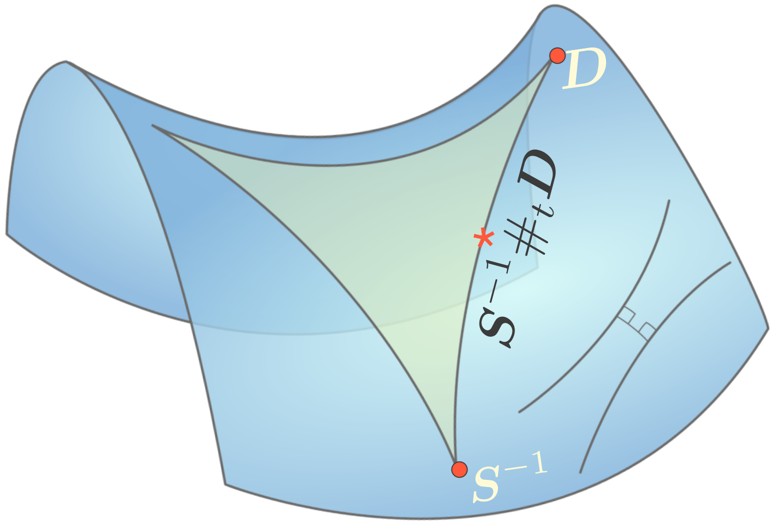
Closed form solution!

$$\nabla h(\boldsymbol{A}) = 0 \quad \Leftrightarrow \quad \boldsymbol{S} - \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} = 0$$

$$\boldsymbol{A} = \boldsymbol{S}^{-1}\#_{\frac{1}{2}}\boldsymbol{D}$$

More generally

$$\min_{\boldsymbol{A}\succ 0} \quad (1-t)\delta_R^2(\boldsymbol{S}^{-1}, \boldsymbol{A}) + t\delta_R^2(\boldsymbol{D}, \boldsymbol{A})$$



$$\boldsymbol{S}^{-1}\#_t\boldsymbol{D}$$

Geometric approach to metric learning

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

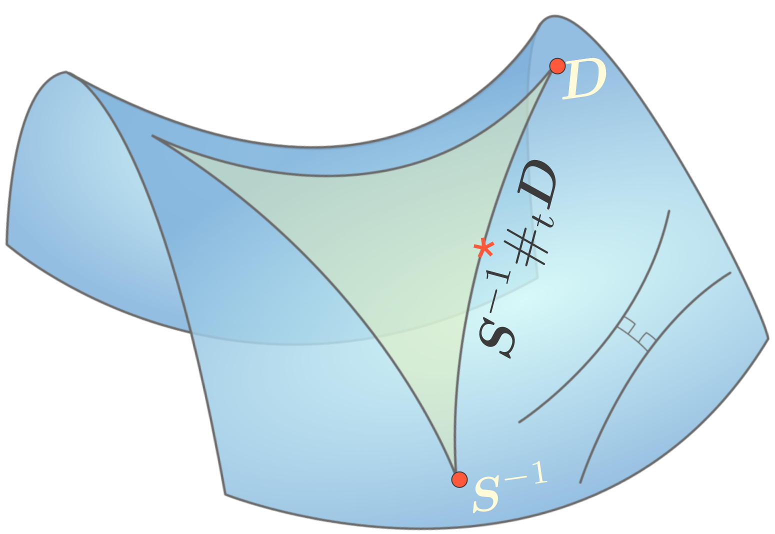
Closed form solution!

$$\nabla h(\boldsymbol{A}) = 0 \quad \Leftrightarrow \quad \boldsymbol{S} - \boldsymbol{A}^{-1}\boldsymbol{D}\boldsymbol{A}^{-1} = 0$$

$$\boldsymbol{A} = \boldsymbol{S}^{-1} \#_{\frac{1}{2}} \boldsymbol{D}$$

More generally

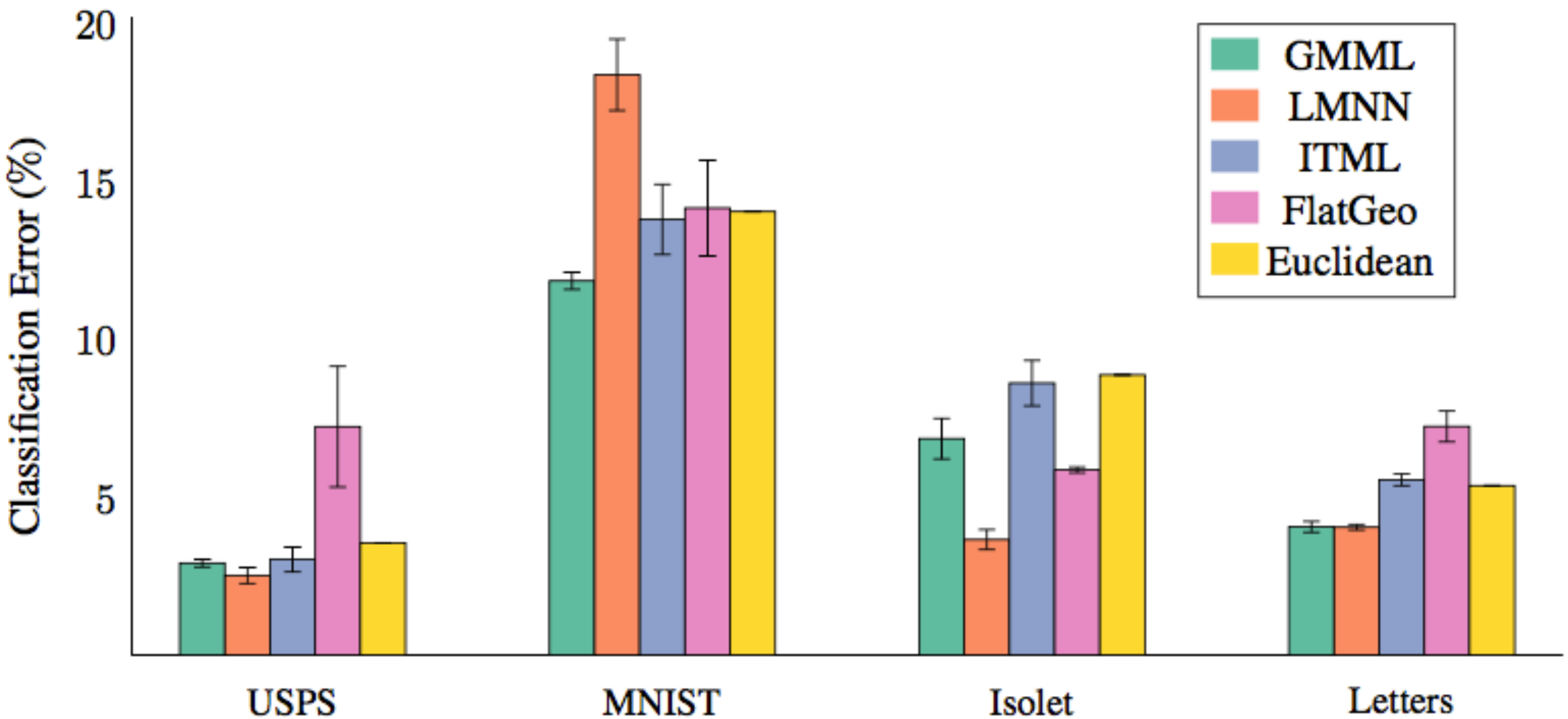
$$\min_{\boldsymbol{A} \succ 0} \quad (1-t)\delta_R^2(\boldsymbol{S}^{-1}, \boldsymbol{A}) + t\delta_R^2(\boldsymbol{D}, \boldsymbol{A})$$



$$\boldsymbol{S}^{-1} \#_t \boldsymbol{D}$$

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

DATA SET	GMML	LMNN	ITML	FLATGEO
SEGMENT	0.0054	77.595	0.511	63.074
LETTERS	0.0137	401.90	7.053	13543
USPS	0.1166	811.2	16.393	17424
ISOLET	1.4021	3331.9	1667.5	24855
MNIST	1.6795	1396.4	1739.4	26640

Comment: May think of this as a "supervised whitening transform"

[Habibzadeh, Hosseini, Sra ICML 2016]

Brascamp-Lieb Constant

Brascamp-Lieb Constant

$$\int_{\mathbb{R}^n} \prod_{i=1}^{m} f_i(B_i x)^{p_i}\, dx \le D^{-1/2} \prod_{i=1}^{m} \left(\int_{\mathbb{R}^{n_i}} f_i(y)\, dy \right)^{p_i}$$

Brascamp-Lieb Constant

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

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

$$p_i > 0, f_i \geq 0 \qquad \sum_{i=1}^m p_i n_i = n$$

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

Important in: Information theory, convex geometry, probability theory

Brascamp-Lieb Constant

$$\int_{\mathbb{R}^n} \prod_{i=1}^m f_i(B_i x)^{p_i}\, dx \leq \boxed{D^{-1/2}} \prod_{i=1}^m \left(\int_{\mathbb{R}^{n_i}} f_i(y)\, dy \right)^{p_i}$$

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

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

$$p_i > 0, f_i \geq 0 \qquad \sum_{i=1}^m p_i n_i = n$$

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

Important in: Information theory, convex geometry, probability theory

Brascamp-Lieb constant

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

- 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

$$\min_{X_1,\ldots,X_m \succ 0} \quad \log\det\Big(\sum\nolimits_i p_i B_i^* X_i B_i\Big) - \sum\nolimits_i p_i \log\det X_i$$

- 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

$$\min_{X_1,\ldots,X_m \succ 0} \log\det\left(\sum\nolimits_i p_i B_i^* X_i B_i\right) - \sum\nolimits_i p_i \log\det X_i$$

- 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

Let's look at a proof….

Proving g-convexity

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

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

Proving g-convexity

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

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

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)

Proving g-convexity

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

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

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. $\Phi(X\#_t Y) \preceq \Phi(X)\#_t\Phi(Y)$.

Proving g-convexity

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

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

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.** $\Phi(X\#_t Y) \preceq \Phi(X)\#_t\Phi(Y)$.
>
> *Proof.* *[Kubo-Ando 1980; Sra-Hosseini 2015]*

Proving g-convexity

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

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

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. $\Phi(X\#_t Y) \preceq \Phi(X)\#_t\Phi(Y)$.

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

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

Proving g-convexity

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

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

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. $\Phi(X\#_t Y) \preceq \Phi(X)\#_t\Phi(Y)$.

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

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

Proof. see e.g., *[Bhatia 2007, "Positive definite matrices"]*

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

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

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

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

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

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

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

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

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

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

$$\sum_i \Phi_i(X_i \#_t Y_i) \preceq \sum_i \Phi_i(X_i) \#_t \Phi_i(Y_i) \preceq \Big(\sum_i \Phi_i(X_i)\Big) \#_t \Big(\sum_i \Phi_i(Y_i)\Big)$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

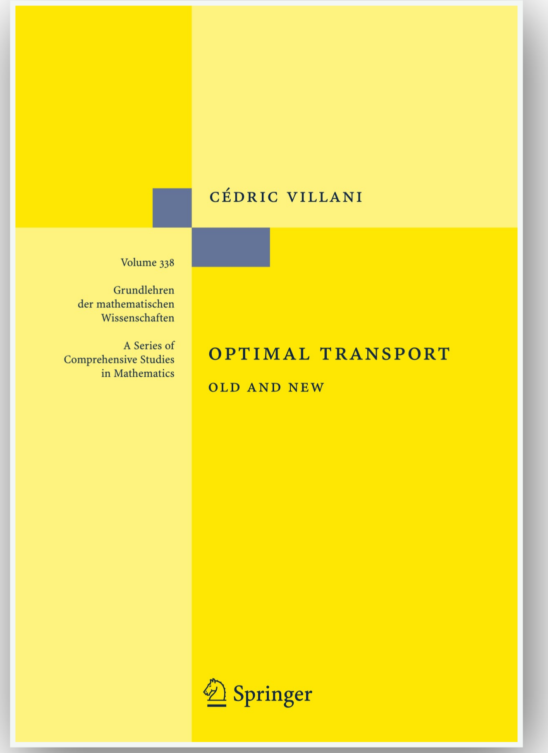
$$\leqslant (1-t)\log\det\left(\sum_i \Phi_i(X_i)\right) + t\log\det\left(\sum_i \Phi_i(Y_i)\right)$$

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

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

Wasserstein barycenters

Wasserstein distance between multivariate Gaussians

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

Wasserstein Barycenter

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

Wasserstein distance between multivariate Gaussians

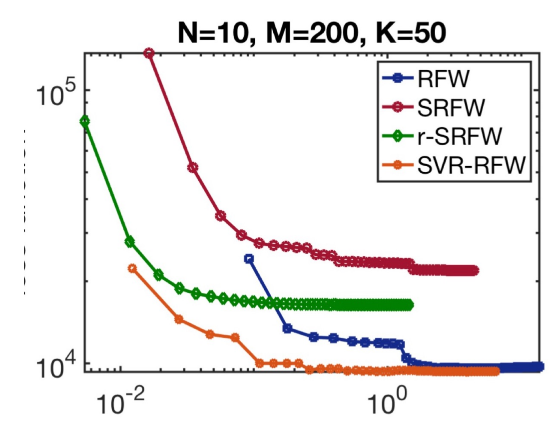
$$d_W(X, Y) = \left[\mathrm{tr}(X+Y) - 2\mathrm{tr}\left(X^{1/2} Y X^{1/2}\right)^{1/2}\right]^{1/2}.$$

Wasserstein Barycenter

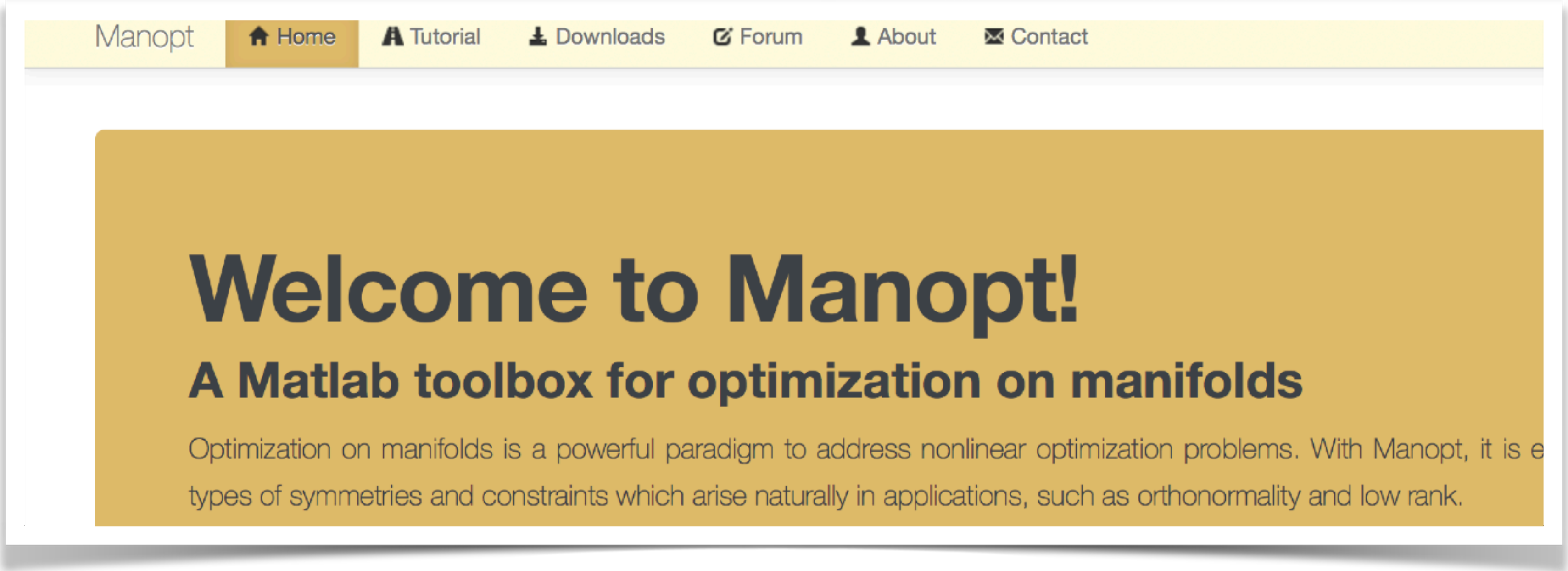
$$\min_{X \succeq 0} \quad \frac{1}{N}\sum_{i=1}^{N} d_W^2(X, A_i)$$

Actually a (Euclidean) convex optimization problem

But empirically Riemannian optimization turns out to be faster!

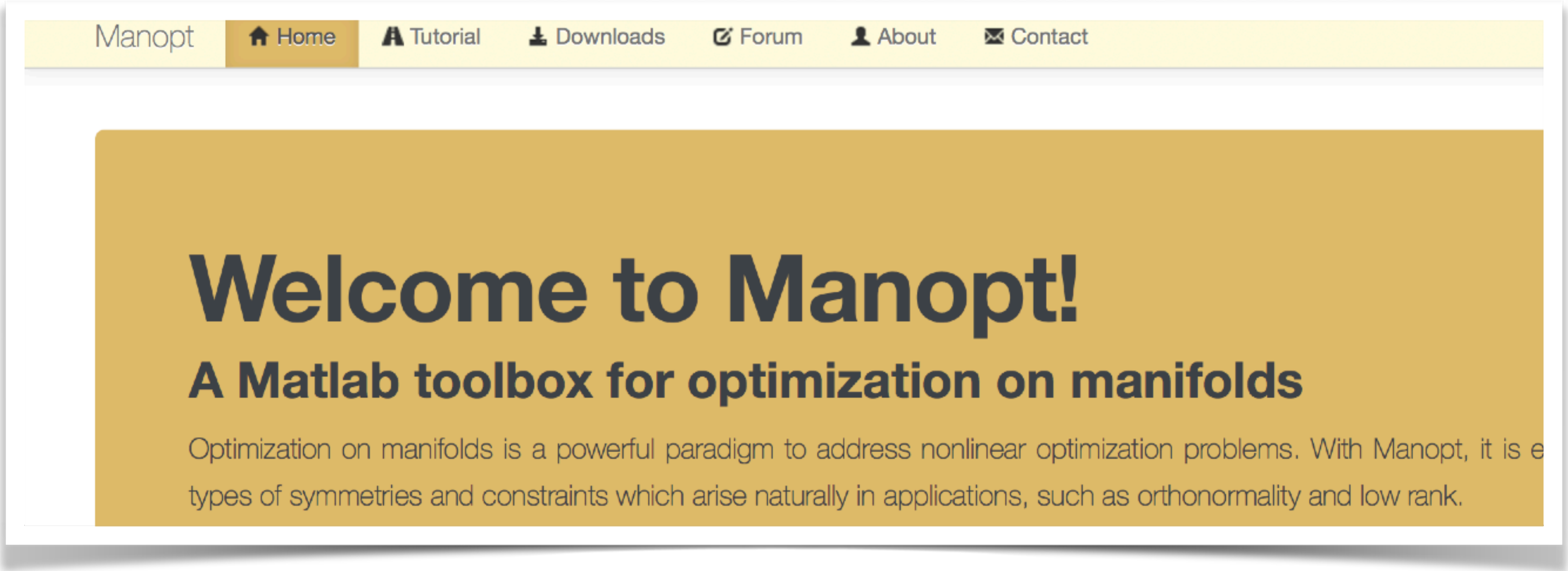


Recent toolboxes, tutorials

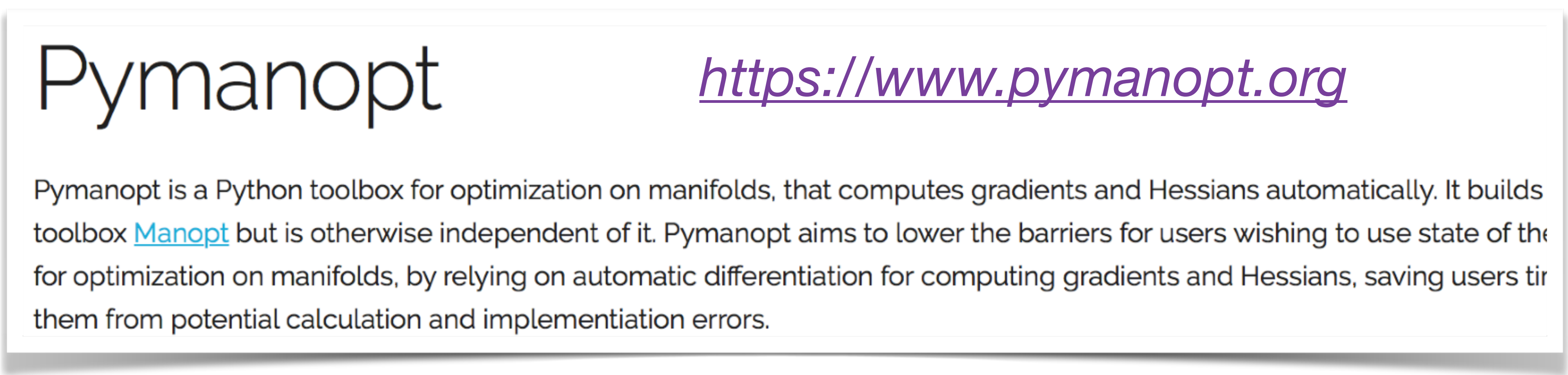


https://www.manopt.org

Recent toolboxes, tutorials



https://www.manopt.org



Recent toolboxes, tutorials

Manopt | Home | Tutorial | Downloads | Forum | About | Contact

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 e
types of symmetries and constraints which arise naturally in applications, such as orthonormality and low rank.

https://www.manopt.org

Pymanopt

https://www.pymanopt.org

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
for optimization on manifolds, by relying on automatic differentiation for computing gradients and Hessians, saving users ti
them from potential calculation and implementiation errors.

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

Recent toolboxes, tutorials

Manopt | Home | Tutorial | Downloads | Forum | About | Contact

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 e
types of symmetries and constraints which arise naturally in applications, such as orthonormality and low rank.

https://www.manopt.org

Pymanopt

https://www.pymanopt.org

Pymanopt is a Python toolbox for optimization on manifolds, that computes gradients and Hes
toolbox Manopt but is otherwise independent of it. Pymanopt aims to lower the barriers for use
for optimization on manifolds, by relying on automatic differentiation for computing gradients a
them from potential calculation and implementiation errors.

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

New book: