GEOMETRIC OPTIMIZATION

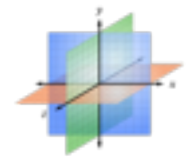
SUVRIT SRA
Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

LIDS Seminar, 13 Sep 2016

Includes work with:
Reshad Hosseini
Pourya H. Zadeh
Hongyi Zhang

π

- **Vector spaces**



- **Manifolds**
  (hypersphere, orthogonal matrices, complicated surfaces)

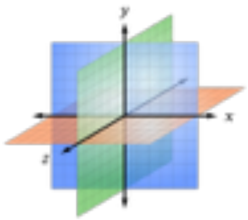
- **Convex sets**
  (probability simplex, semidefinite cone, polyhedra)

- **Metric spaces**
  (tree space, Wasserstein spaces, CAT(0), space-of-spaces)

Geometric Optimization

Machine Learning
Graphics
Robotics
Vision
BCI
NLP
Statistics

Example: Riemannian optimization



Vector space optimization

Orthogonality constraint

Fixed-rank constraint

Positive semi-definite constraint

... ...

⇔

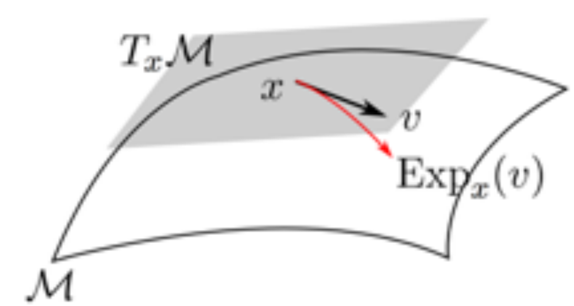
Stiefel manifold

Grassmann manifold

PSD manifold

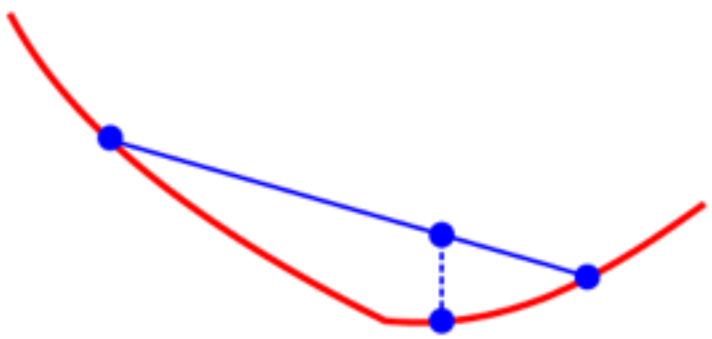
... ...

Riemannian optimization

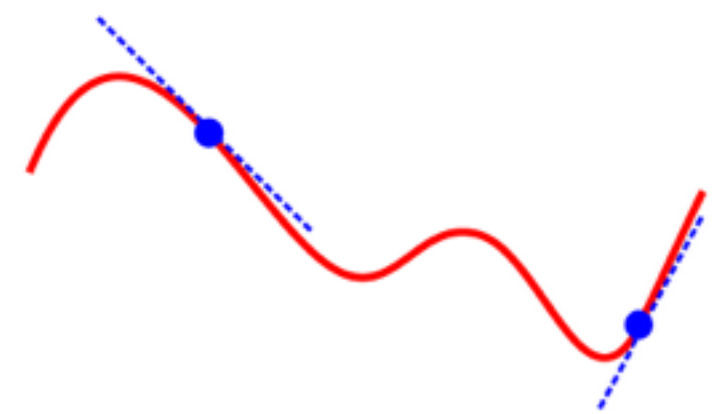


[Udriste, 1994; Absil et al., 2009]

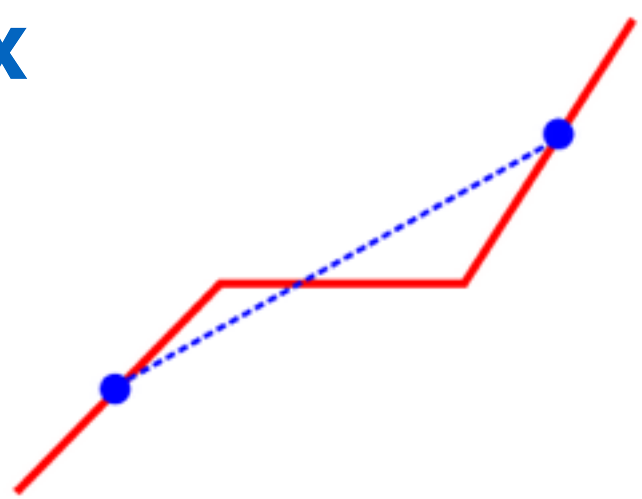
Function classes of interest



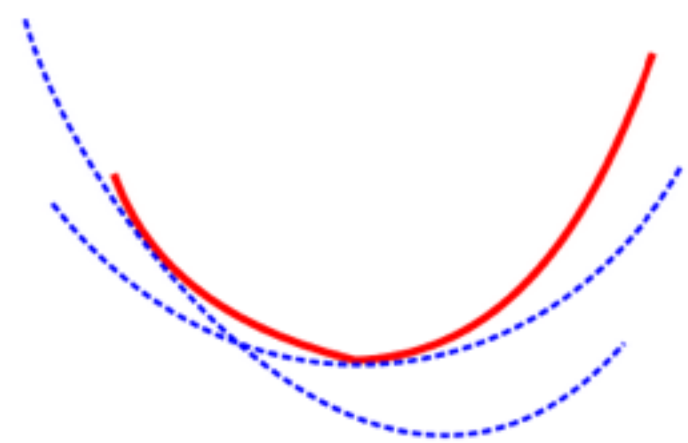
Convex



Smooth

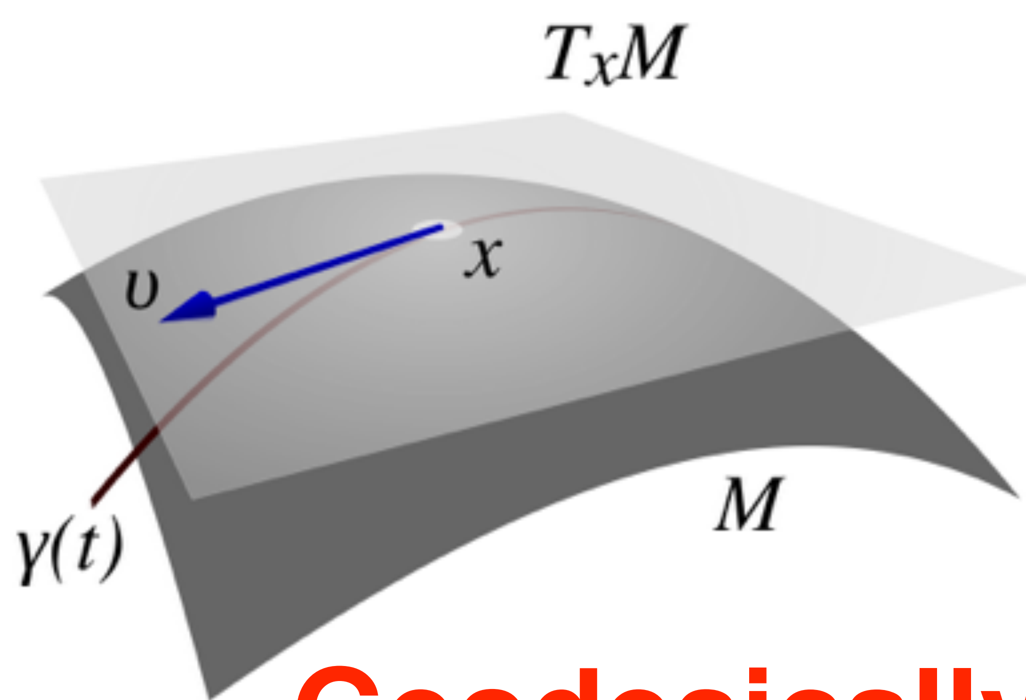


Lipschitz

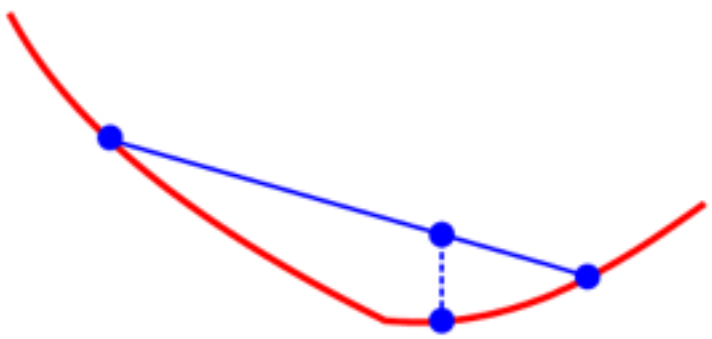


Strongly convex

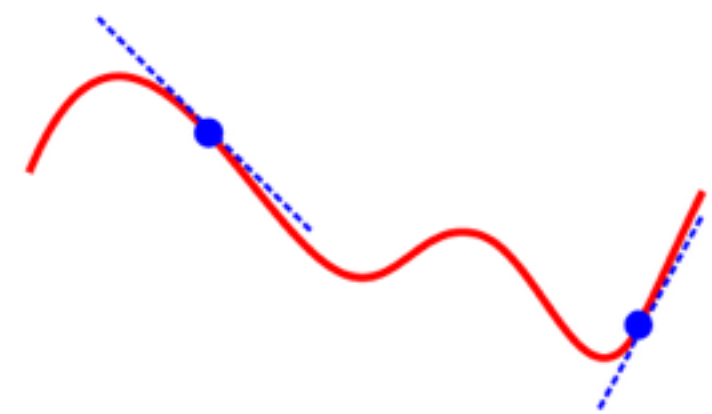
Function classes of interest



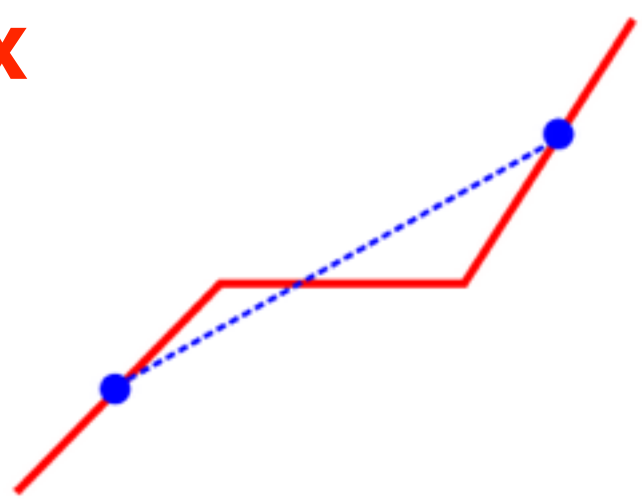
Geodesically



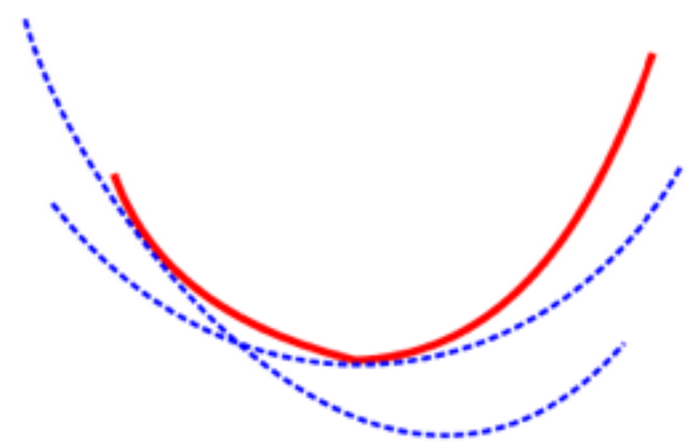
Convex



Smooth



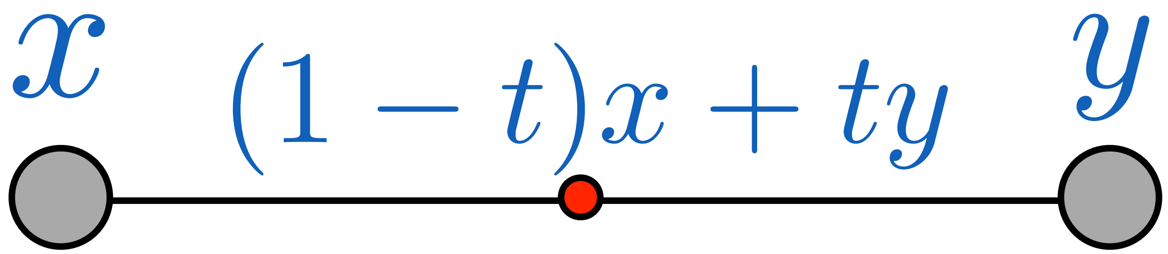
Lipschitz



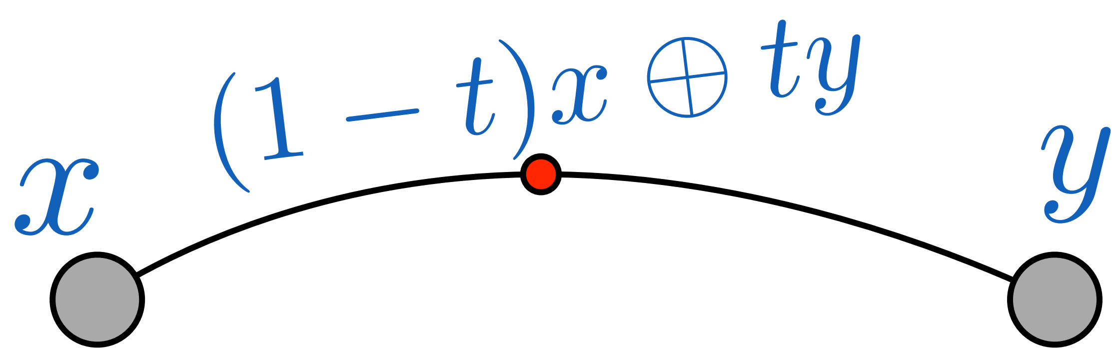
Strongly convex

What is geodesic convexity?

Convexity



Geodesic convexity



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

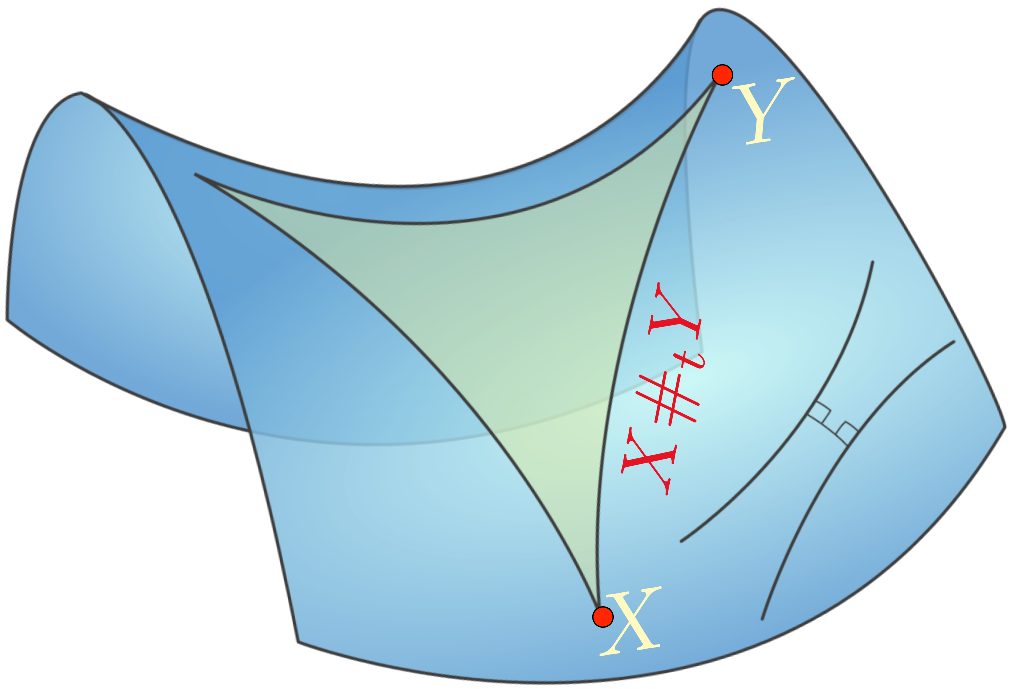
on riemannian manifold $f(y) \geq f(x) + \langle g_x, \mathrm{Exp}_x^{-1}(y) \rangle_x$

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

Positive definite matrix manifold

Geodesic

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



Examples

$$f(X) = \begin{cases} \log\det(X), & \log\operatorname{tr}(X), \\ \operatorname{tr}(X^\alpha), & \|X^\alpha\|. \end{cases}$$

Verify

$$f(X\#_t Y) \le (1-t)f(X) + tf(Y)$$

Positive definite matrix manifold

Recognizing, constructing, and optimizing g-convex functions

[Sra, Hosseini (2013,2015)]

– *[Wiesel 2012]*
– *[Rápcsák 1984]*
– *[Udriste 1994]*

Corollaries

$$X \mapsto \log\det(B + \sum_i A_i^* X A_i)$$

$$X \mapsto \log\mathrm{per}(B + \sum_i A_i^* X A_i)$$

$$\delta_R^2(X, Y), \quad \delta_S^2(X, Y)$$

(jointly g-convex)

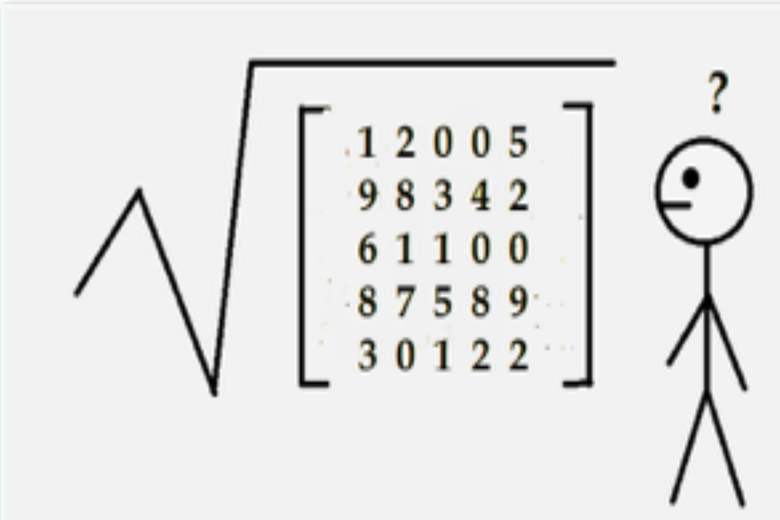
Many more theorems and corollaries

One-D version known as: **Geometric Programming**
www.stanford.edu/~boyd/papers/gp_tutorial.html
[Boyd, Kim, Vandenberghe, Hassibi (2007). 61pp.]

Examples

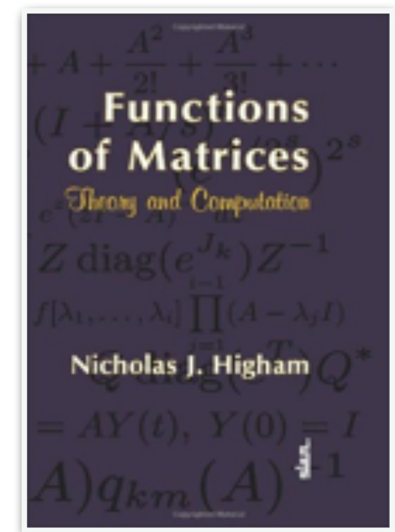
$$X \succ 0$$

Matrix square root



Broadly applicable

Key to ‘`expm`’, ‘`logm`’



Matrix square root

Nonconvex optimization through the Euclidean lens

[Jain, Jin, Kakade, Netrapalli; Jul 2015]

$$\min_{X \in \mathbb{R}^{n \times n}} \|M - X^2\|_F^2$$

Gradient descent

$$X_{t+1} \leftarrow X_t - \eta(X_t^2 - M)X_t - \eta X_t(X_t^2 - M)$$

Simple algorithm; linear convergence; **nontrivial** analysis

Matrix square root

Geodesic

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

Midpoint

$$A^{\frac{1}{2}} = A \#_{\frac{1}{2}} I$$

Matrix square root

Nonconvex optimization through **non-Euclidean** lens

[Sra; Jul 2015]

$$\min_{X \succ 0} \quad \delta_S^2(X, A) + \delta_S^2(X, I)$$

Fixed-point iteration

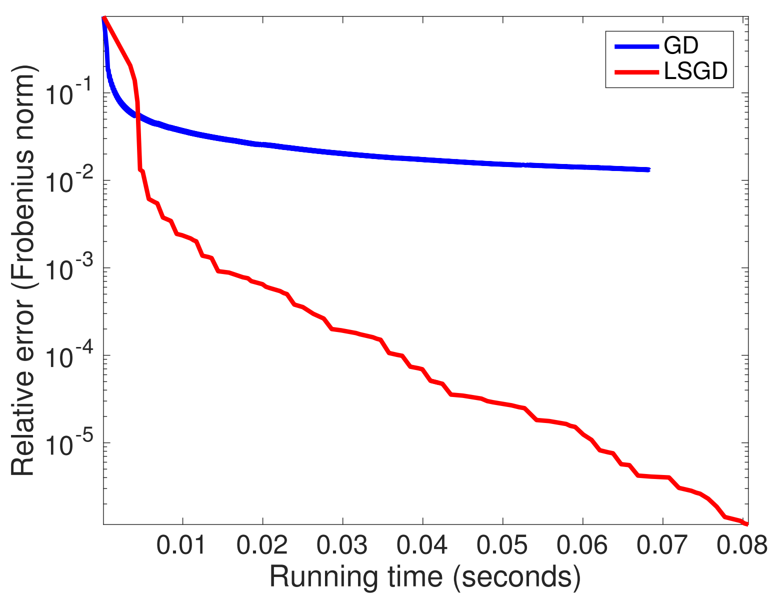
$$X_{k+1} \leftarrow [(X_k + A)^{-1} + (X_k + I)^{-1}]^{-1}$$

Simple method; linear convergence; 1/2 page analysis!

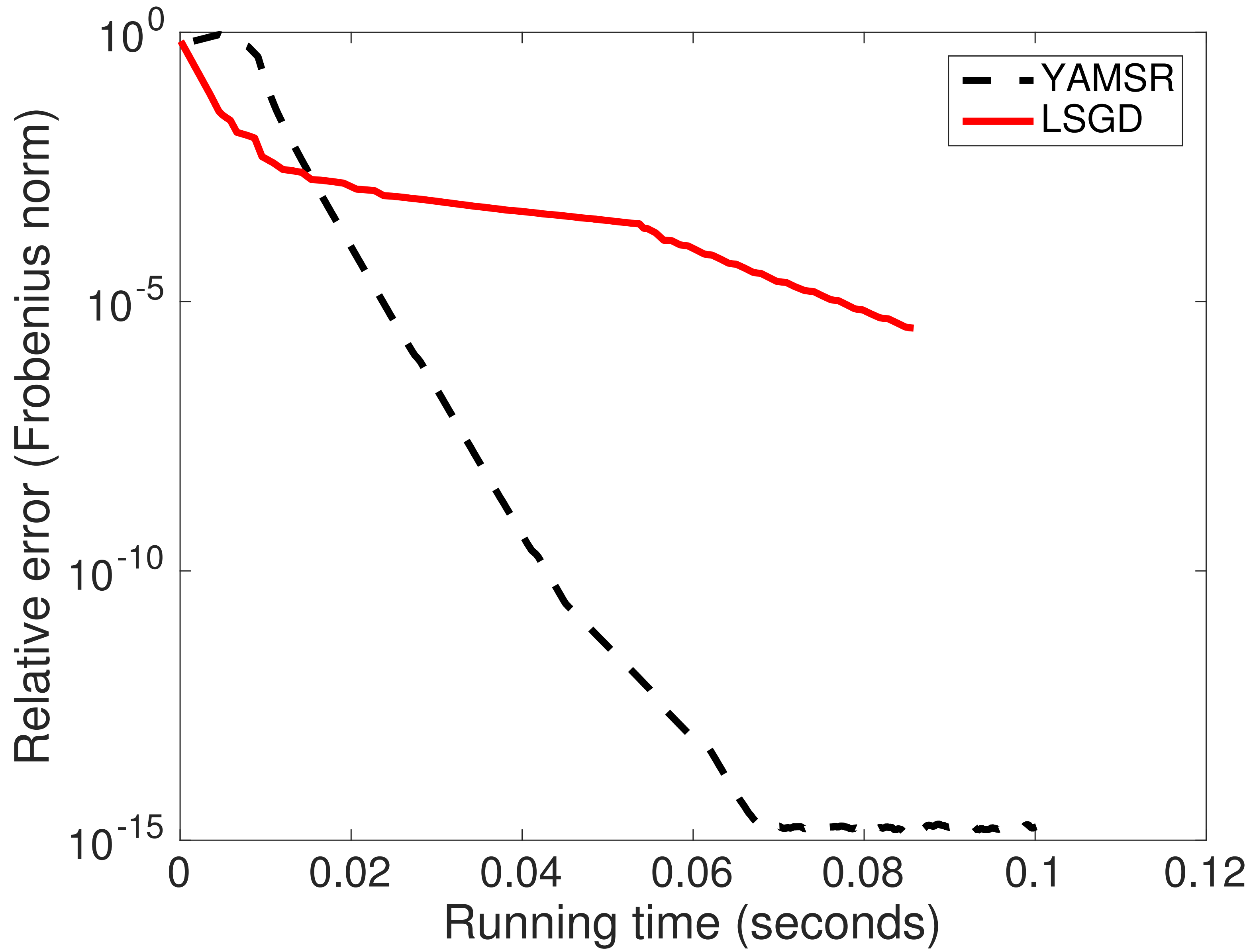
Global optimality thanks to geodesic convexity

$$\delta_S^2(X, Y) := \tfrac{1}{2} \log\det\left(\tfrac{X+Y}{2}\right) - \tfrac{1}{2}\log\det(XY)$$

Matrix square root



$$50 \times 50 \text{ matrix } I + \beta UU^T$$
$$\kappa \approx 64$$



Brascamp-Lieb Constant

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

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

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

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

Brascamp-Lieb constant

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

- Solved and analyzed via an elaborate approach in:
  [Garg, Gurvits, Oliveira, Wigderson; Jul 2016]

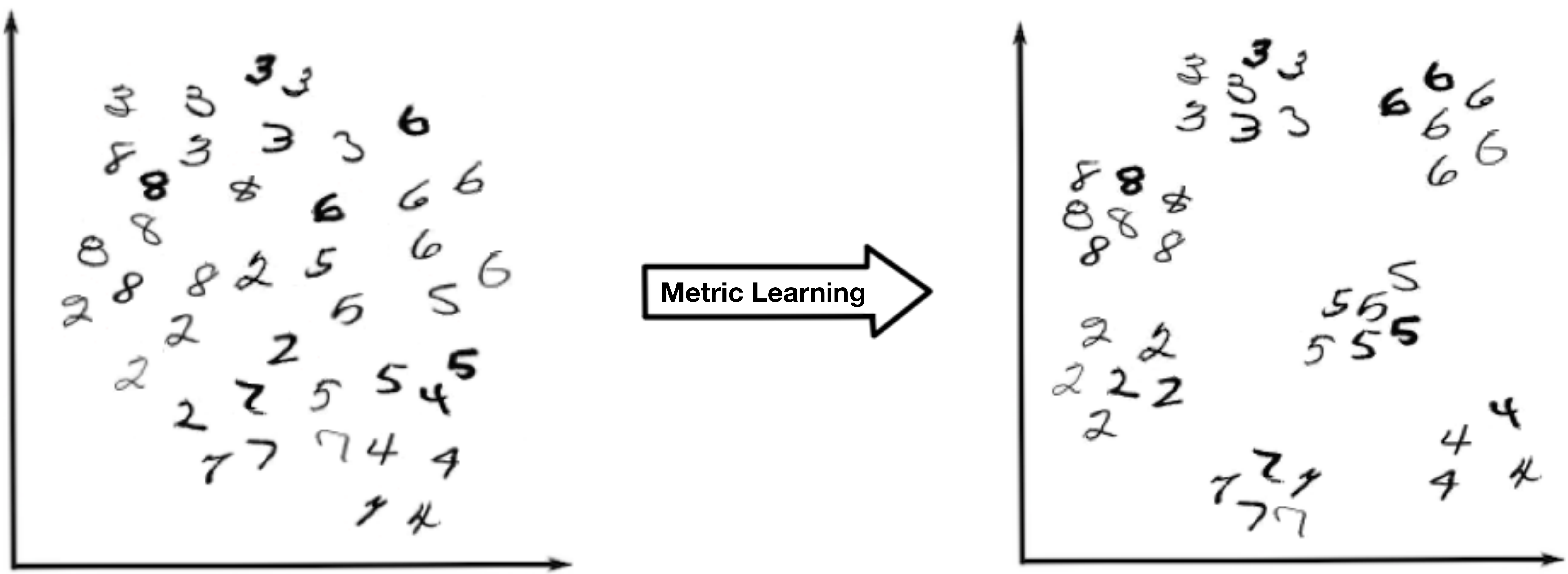
Exercise

Prove this is a g-convex opt problem

- G-convexity yields transparent algorithms & complexity analysis for global optimum!

Metric learning

What does a metric learning method do?



[Habibzadeh, Hosseini, Sra, ICML 2016]

Euclidean metric learning

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

given pairwise constraints learn Mahalanobis distance

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

Positive definite matrix ***A***

Metric learning methods

MMC

[Xing, Jordan, Russell, Ng 2002]

$$\min_{A \succeq 0} \sum_{(x_i, x_j) \in \mathcal{S}} d_A(x_i, x_j)$$
$$\text{such that} \sum_{(x_i, x_j) \in \mathcal{D}} \sqrt{d_A(x_i, x_j)} \geq 1$$

LMNN

[Weinberger, Saul 2005]

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

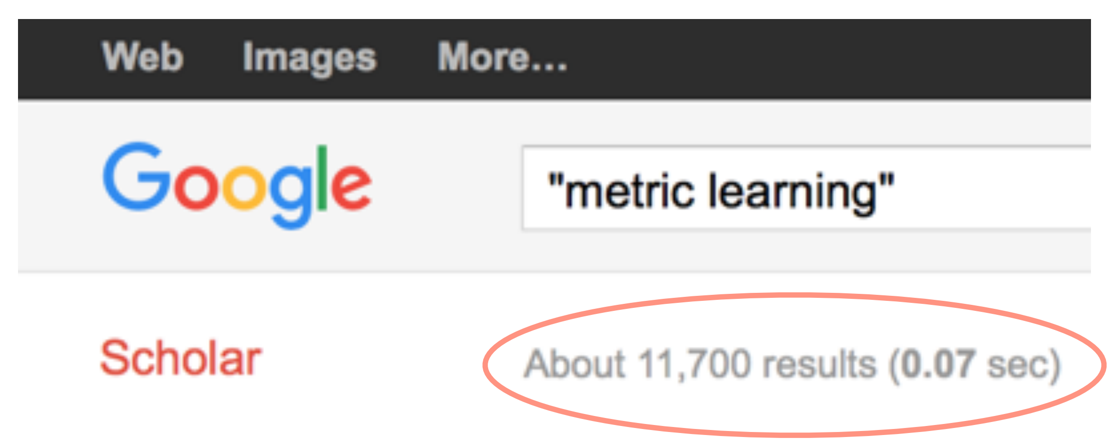
ITML

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

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

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

tons of other methods!



A simple new way for metric learning

Euclidean idea

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

New idea

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

Equivalently solve

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

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

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

A simple new way for metric learning

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

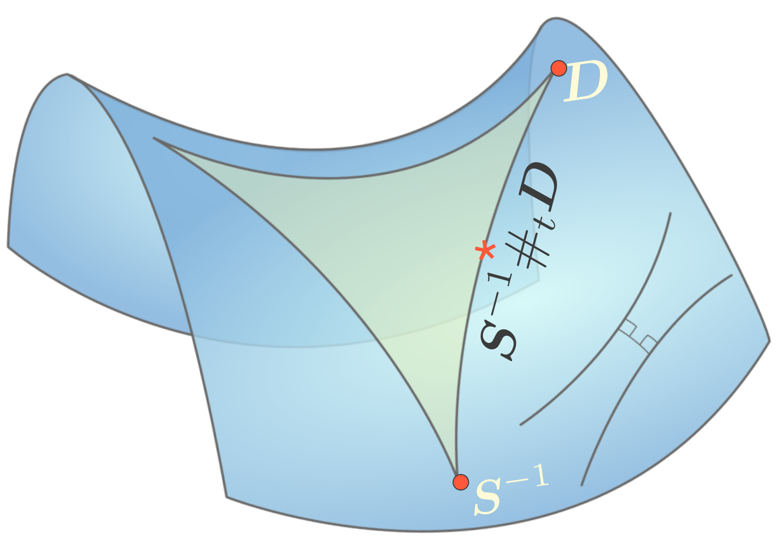
Closed form solution!

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

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

More generally

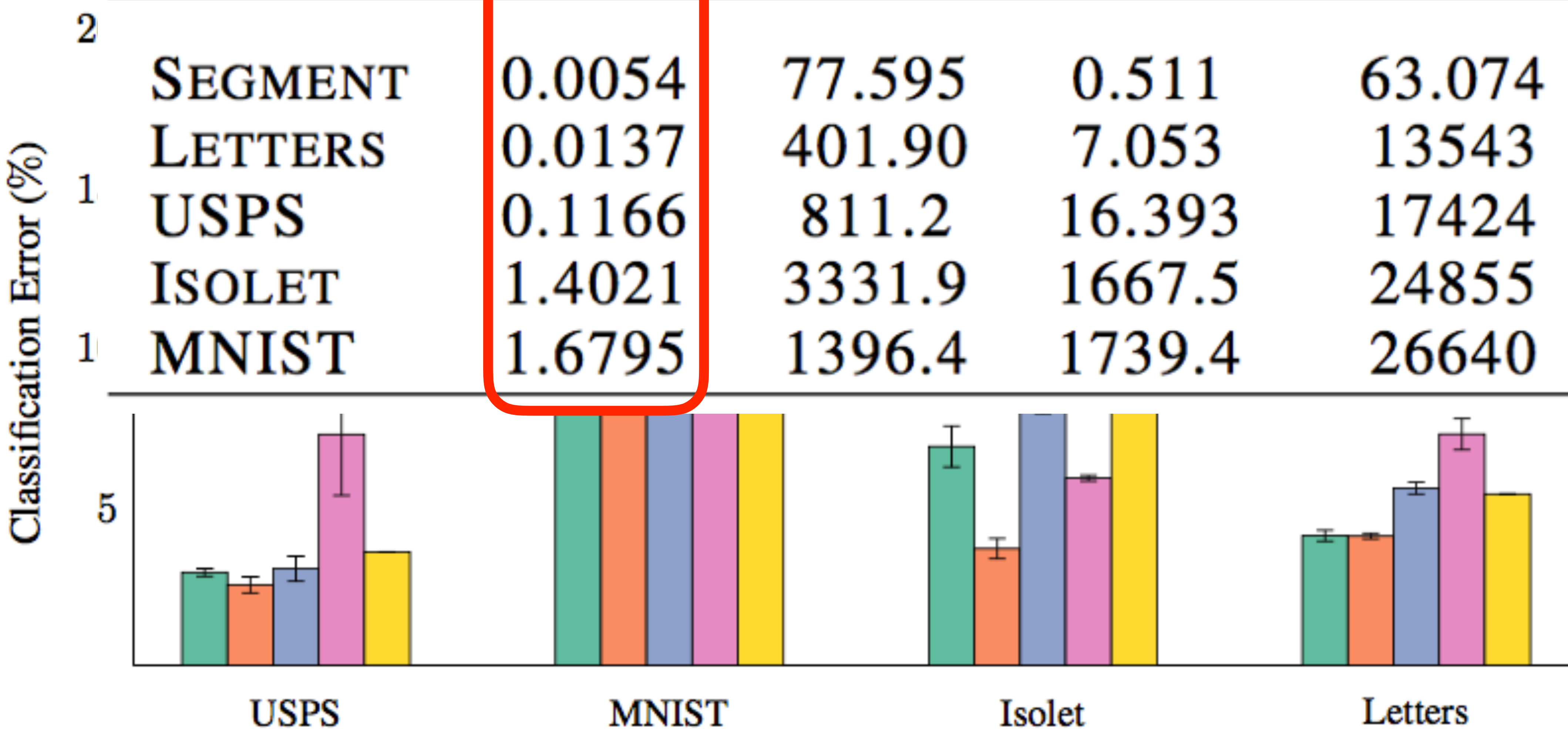
$$\min_{\boldsymbol{A} \succ 0} \quad (1-t)\delta_R^2(\boldsymbol{S}^{-1}, \boldsymbol{A}) + t\delta_R^2(\boldsymbol{D}, \boldsymbol{A})$$



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

Experiments

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

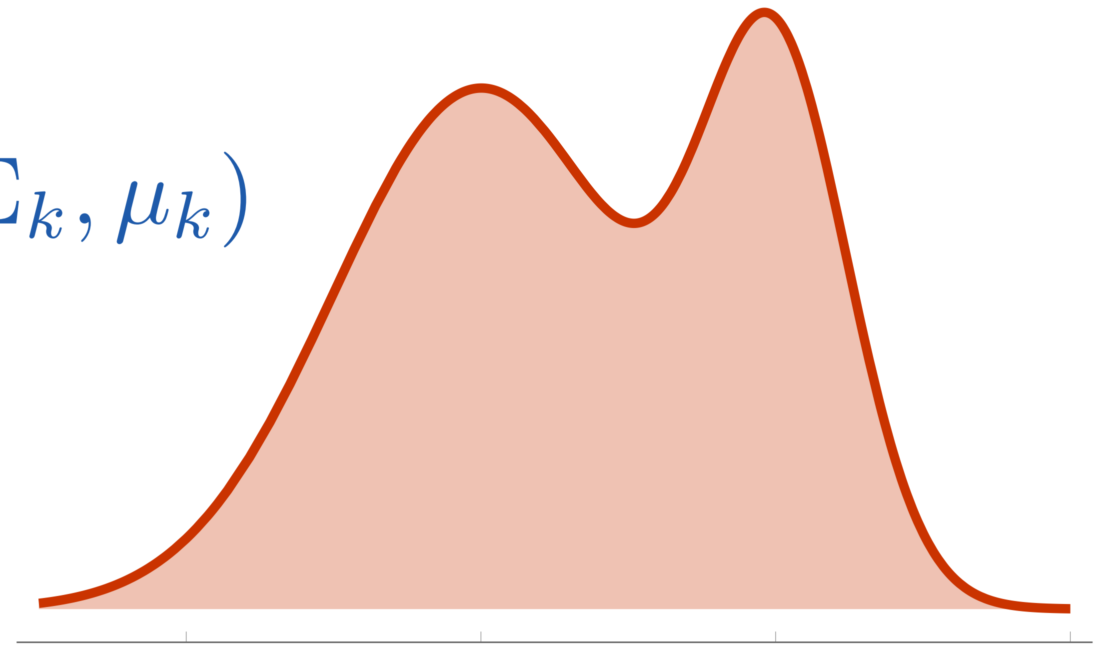


[Habibzadeh, Hosseini, Sra ICML 2016]

Gaussian mixture models

$$p_{\text{mix}}(x) := \sum_{k=1}^{K} \pi_k p_{\mathcal{N}}(x; \Sigma_k, \mu_k)$$

$$\max \prod_i p_{\text{mix}}(x_i)$$



Expectation maximization (EM): default choice

$$p_{\mathcal{N}}(x; \Sigma, \mu) \propto \frac{1}{\sqrt{\det(\Sigma)}} \exp\left(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)\right)$$

[Hosseini, Sra NIPS 2015]

Gaussian mixture models

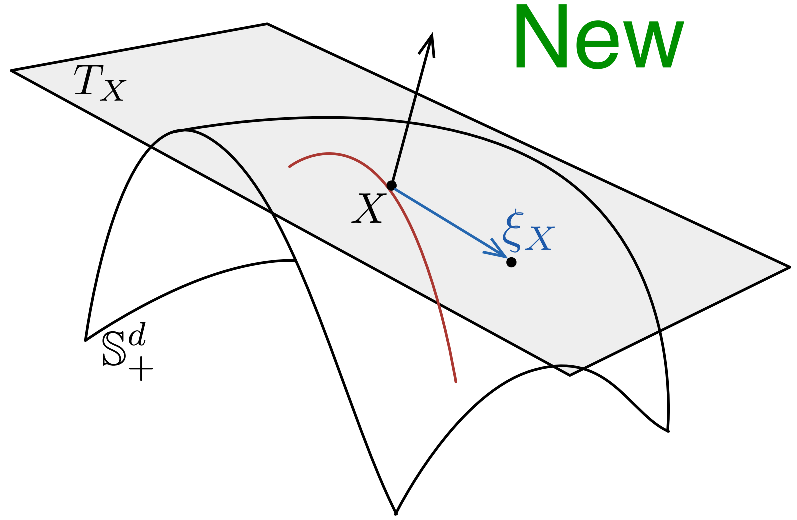
- **Nonconvex –** difficult, possibly several local optima
- **GMMs –** Recent surge of theoretical results
- **In Practice –** EM still default choice

(Often claimed that standard nonlinear programming algorithms inferior for GMMs)

Difficulty: Positive definiteness constraint on Σ_k

Geometric opt

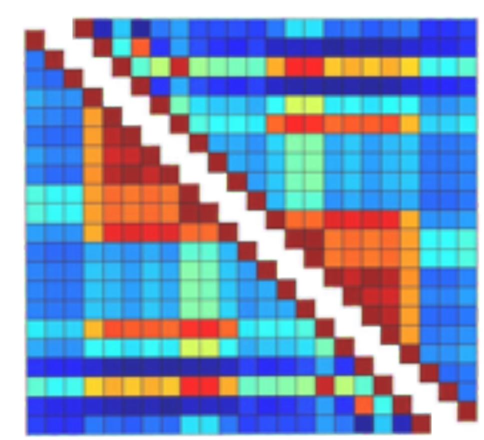
New



Unconstrained, Cholesky

Folklore

LL^T



Failure of geometric optimization

K	EM	Riem-CG
2	17s // 29.28	947s // 29.28
5	202s // 32.07	5262s // 32.07
10	2159s // 33.05	17712s // 33.03

manopt.org

Riemannian opt. toolbox

d=35
images
dataset

Failure of “obvious” LL^T

sep.	EM	CG-LL^T
0.2	52s // 12.7	614s // 12.7
1	160s // 13.4	435s // 13.5
5	72s // 12.8	426s // 12.8

$$\|\mu_i - \mu_j\| \geq \text{ sep } \max_{ij}\{\text{tr}\Sigma_i, \text{tr}\Sigma_j\}$$

d=20
simulation

What's wrong?

log-likelihood for one component

$$\max_{\mu,\Sigma\succ 0} \mathcal{L}(\mu,\Sigma) := \sum_{i=1}^{n} \log p_{\mathcal{N}}(x_i;\mu,\Sigma).$$

$$-\frac{n}{2}\log\det\Sigma - \frac{1}{2}\sum\nolimits_{i=1}^{n}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

Euclidean convex problem
Not geodesically convex

Mismatched geometry?

Reformulate as g-convex

$$y_i = \begin{bmatrix} x_i \\ 1 \end{bmatrix} \quad S = \begin{bmatrix} \Sigma + \mu\mu^T & \mu \\ \mu^T & 1 \end{bmatrix}$$

$$\max_{S \succ 0} \; \widehat{\mathcal{L}}(S) := \sum_{i=1}^{n} \log q_{\mathcal{N}}(y_i; S),$$

Thm. The modified log-likelihood is g-convex. Local max of modified mixture LL is local max of original.

Success of geometric optimization

K	EM	Riem-CG	L-RBFGS
2	17s // 29.28	**18s** // 29.28	**14s** // 29.28
5	202s // 32.07	**140s** // 32.07	**117s** // 32.07
10	2159s // 33.05	**1048s** // 33.06	**658s** // 33.06

Riem-CG (manopt) savings:

947→**18**; 5262 →**140**; 17712→**1048**

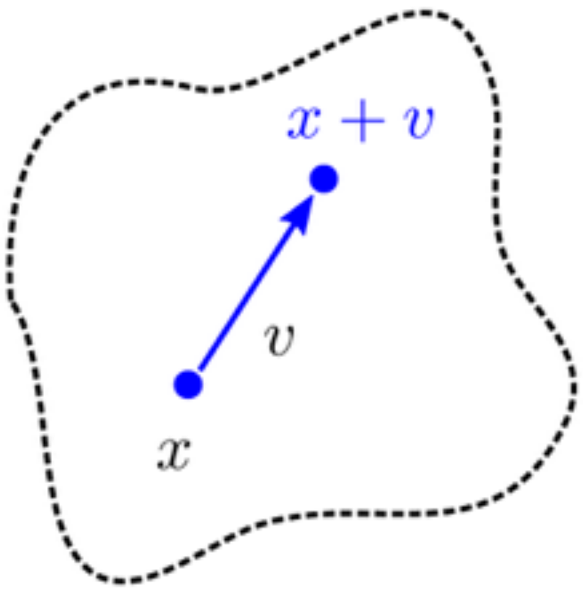
d=35
images
dataset

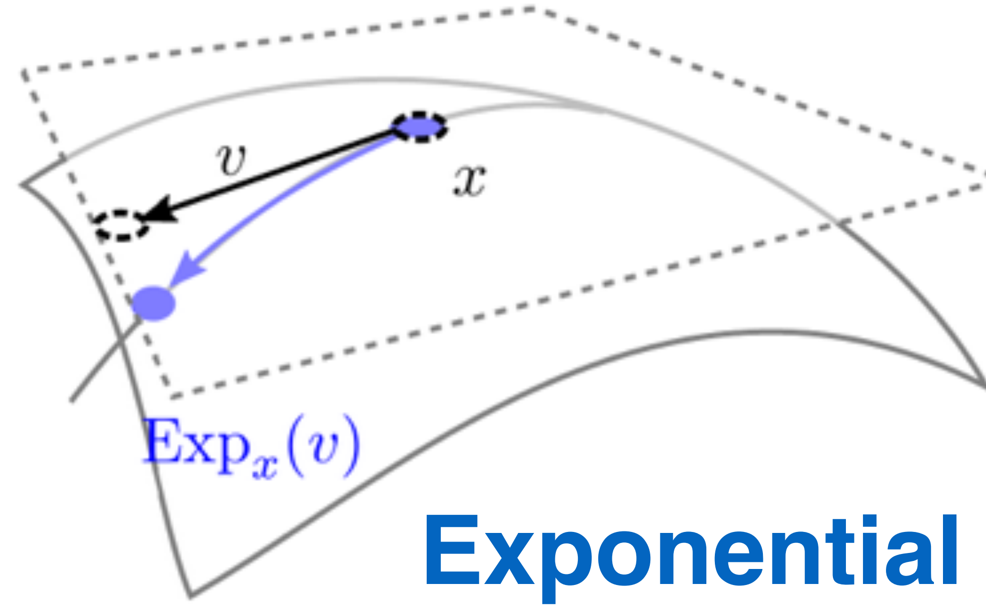
github.com/utvisionlab/mixest

First-order algorithms

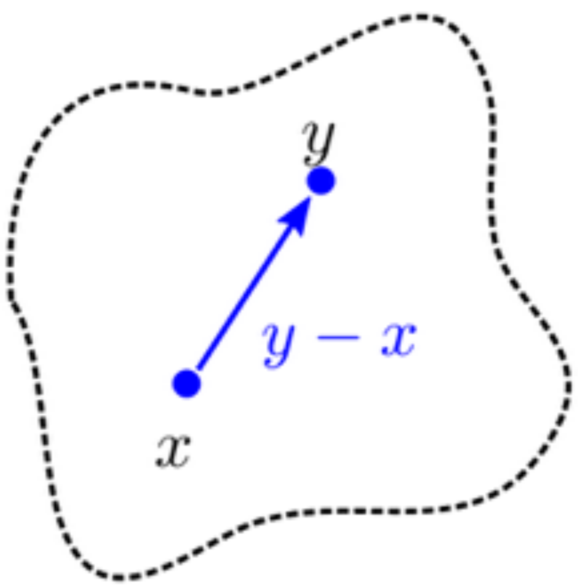
[Zhang, Sra, COLT 2016]

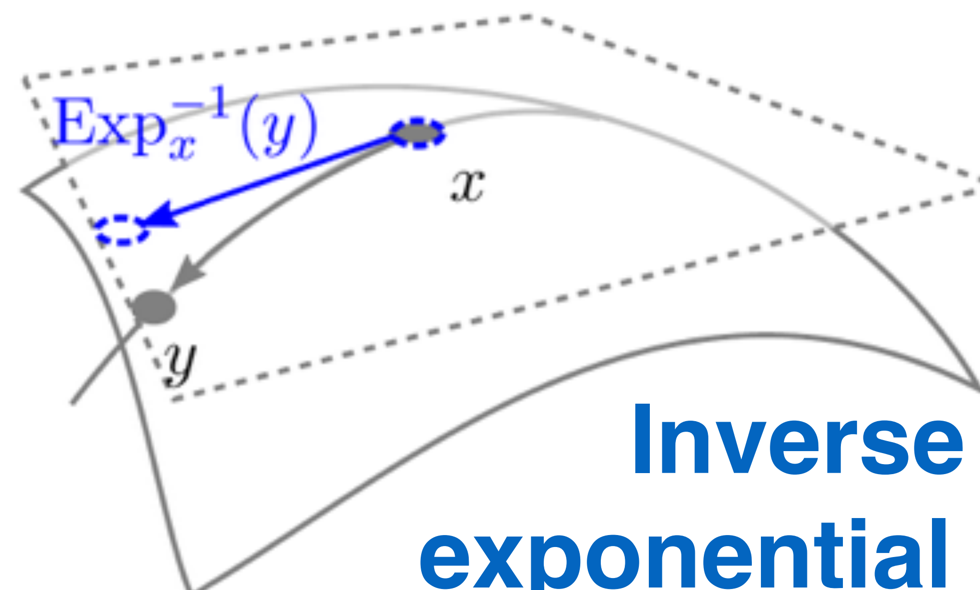
Key concepts generalize





Exponential map





Inverse exponential map

lengths, angles, differentiation, vector translation, etc.

first-order g-convex optimization

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

$\mathcal{X}$ g-convex set; f g-convex func; $\mathcal{M}$ Riemannian manifold

oracle access to exact or stochastic (sub)gradients

$$x \leftarrow \mathrm{Exp}_x(-\eta \nabla f(x))$$

analog to: $x \leftarrow x - \eta \nabla f(x)$

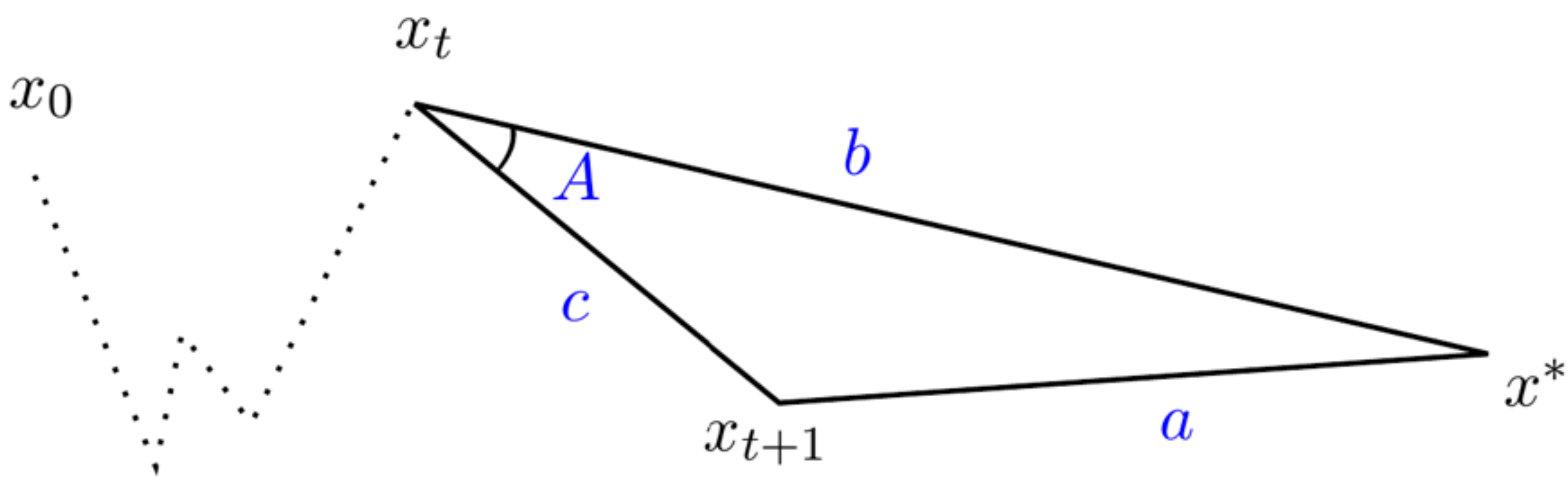
In particular, we study the global complexity of first-order g-convex optimization

Global Complexity

Convex Optimization	G-Convex Optimization
Gradient Descent Stochastic Gradient Descent Coordinate Descent Accelerated Gradient Descent Fast Incremental Gradient	$\mathbb{E}[f(x_a) - f(x^*)] \leq ?$

The Euclidean law of cosines is essential to bound $d^2(x_{t+1},x^*)$ in analysis of usual convex opt. methods

$$x_{t+1} = x_t - \eta_t g_t$$



$$a^2 = b^2 + c^2 - 2bc\cos(A)$$

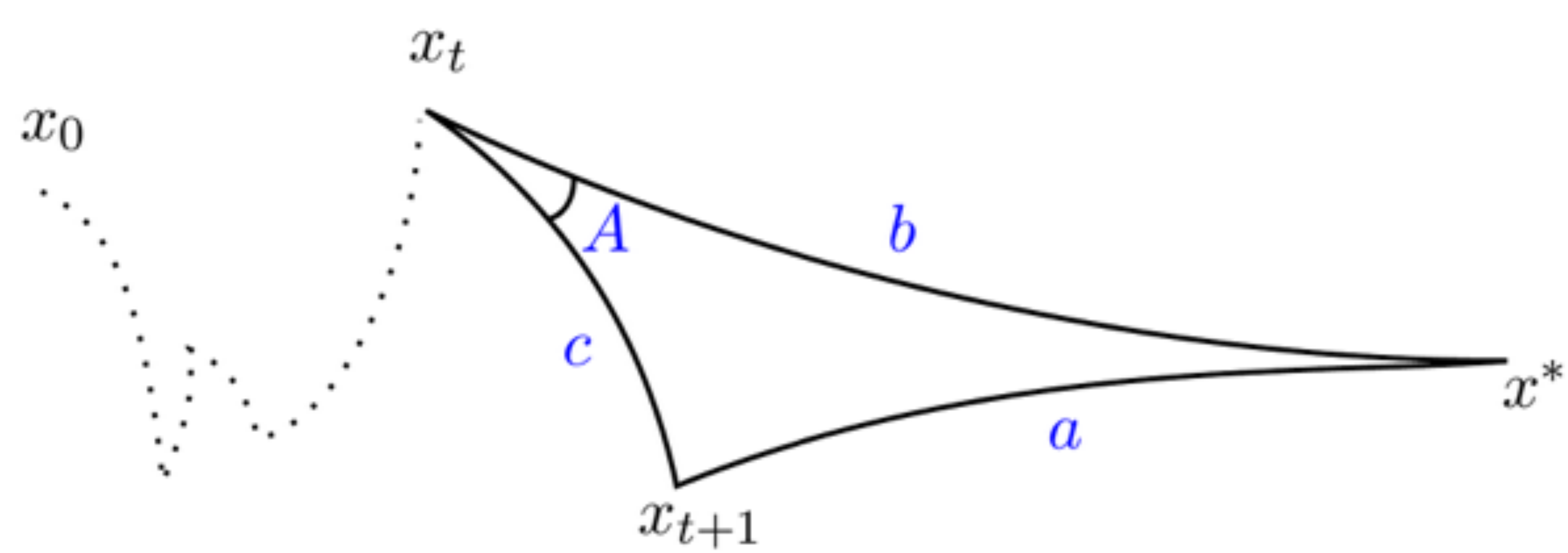
$$\|x_{t+1} - x^*\|^2 = \|x_t - x^*\|^2 + \eta_t^2\|g_t\|^2 - 2\eta_t \langle g_t, x_t - x^* \rangle$$

We develop a corresponding inequality to bound $d^2(x_{t+1}, x^*)$ on manifolds

$$\cosh(-\kappa a) = \cosh(-\kappa b)\cosh(-\kappa c) + \sinh(-\kappa b)\sinh(-\kappa c)\cos(A)$$

Toponogov's theorem

Grönwall's inequality



$$a^2 \le b^2 + \zeta(\kappa_{\min}, b)c^2 - 2bc\cos(A)$$

$$\zeta(\kappa_{\min}, b) \triangleq \frac{\sqrt{|\kappa_{\min}|}b}{\tanh(\sqrt{|\kappa_{\min}|}b)}$$

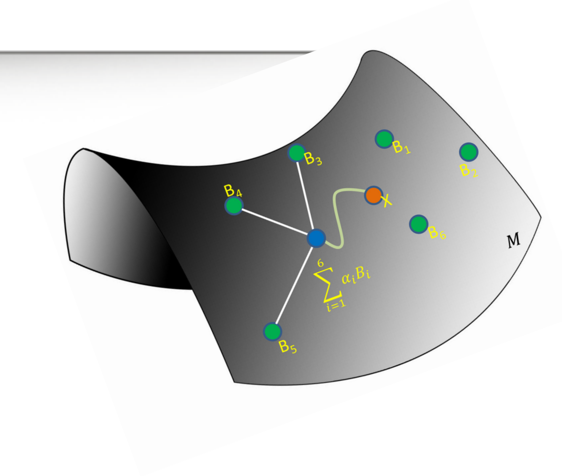
Convergence rates depend on lower bounds on the sectional curvature

	convex	g-convex
(Sub)gradient		
Lipschitz	$O\left(\sqrt{\frac{1}{t}}\right)$	$O\left(\sqrt{\frac{\zeta_{\max}}{t}}\right)$
Strongly convex / smooth	$O\left(\frac{1}{t}\right)$	$O\left(\frac{\zeta_{\max}}{t}\right)$
Strongly convex & smooth	$O\left((1-\frac{\mu}{L_g})^t\right)$	$O\left((1-\min\{\frac{1}{\zeta_{\max}}, \frac{\mu}{L_g}\})^t\right)$
Stochastic (sub)gradient	... ...	

$$\zeta_{\max} \triangleq \frac{\sqrt{|\kappa_{\min}|}D}{\tanh\left(\sqrt{|\kappa_{\min}|}D\right)}$$

See paper for other interesting results *[Zhang, Sra, COLT 2016]*

G-nonconvex optimization



$$\min_{x \in \mathcal{M}} \quad f(x) := \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

- $\mathcal{M}$ is a Riemannian manifold
- g-convex and g-nonconvex 'f' allowed!
- First global complexity results for stochastic methods on general Riemannian manifolds
- Can be faster than Riemannian SGD
- New insights into eigenvector computation

[Zhang, Reddi, Sra, NIPS 2016]