Geometric nonconvex optimization

SUVRIT SRA

Laboratory for Information and Decision Systems
Massachusetts Institute of Technology

TRIPODS MADISON WORKSHOP 2018
July 31st 2018

ml.mit.edu

Key directions for non-convexity in ML

Two main directions

Large-scale nonconvex

Theory & models

Key directions for non-convexity in ML

Two main directions

Large-scale nonconvex

Neural nets, saddle points
Beyond SGD, local min

Theory & models

Bach, Sra (2016). Tutorial at NIPS 2016

"Beyond Stochastic Gradient Descent and Convexity"

[Reddi, Sra, Poczos, Smola, 2018; 2017; 2016a,b,c,d]

[Yun, Sra, Jadbabaie, 2017; 2018]

Key directions for non-convexity in ML

Two main directions

Large-scale nonconvex

Neural nets, saddle points
Beyond SGD, local min

Theory & models

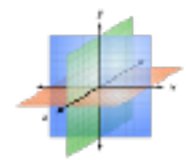
Global optimality via geometry, new ideas, ML models, surprises

Bach, Sra (2016). Tutorial at NIPS 2016
"Beyond Stochastic Gradient Descent and Convexity"

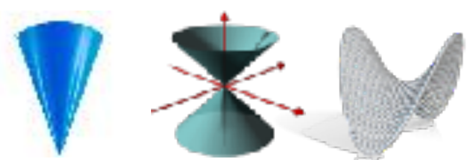
[Reddi, Sra, Poczos, Smola, 2018; 2017; 2016a,b,c,d]
[Yun, Sra, Jadbabaie, 2017; 2018]

What do I mean by Geometry?

- **Vector spaces**
  (the usual setting)



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



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

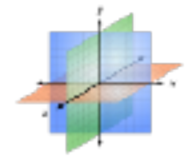
- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

Machine Learning
Graphics
Robotics
Control
Vision
BCI
NLP
Statistics
Biology
and more…

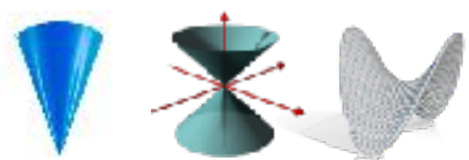
Aim: Use geometry to address non-convex problems

What do I mean by Geometry?

- **Vector spaces**
  (the usual setting)



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



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

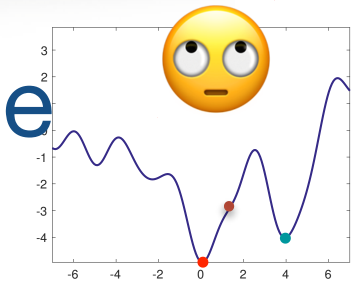
- **Metric spaces**
  (tree space, Wasserstein spaces, space-of-spaces)

Machine Learning
Graphics
Robotics
Control
Vision
BCI
NLP
Statistics
Biology
and more…

Aim: Use geometry to address non-convex problems

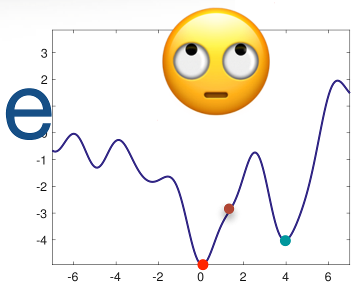
In pursuit of global optimality

Fact: In general, non-convex problems are intractable

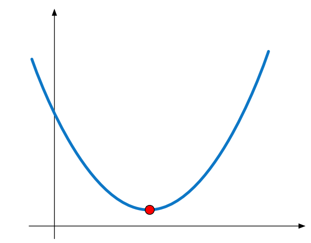


In pursuit of global optimality

Fact: In general, non-convex problems are intractable



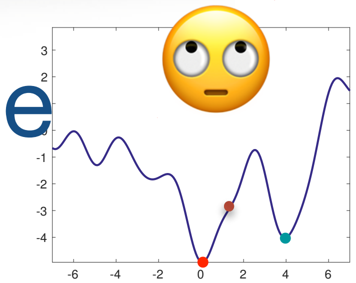
Nature makes exception for convex



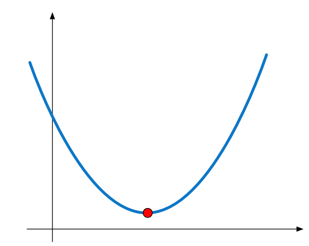
Question: Are convex functions the only exception?

In pursuit of global optimality

Fact: In general, non-convex problems are intractable



Nature makes exception for convex

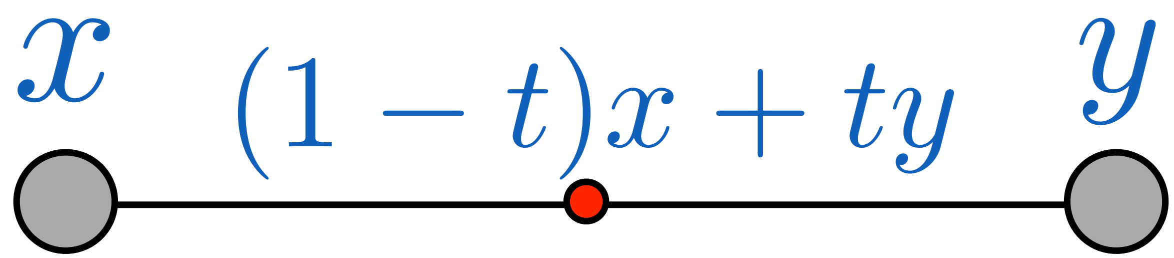


Question: Are convex functions the only exception?

> **Informally** *(Rapcsák, Csendes, 1993)*: If on a nice set A, a function f satisfies local optimum is global optimum, we can reparametrize f to be geodesically convex.

The idea of geodesic convexity

Convexity

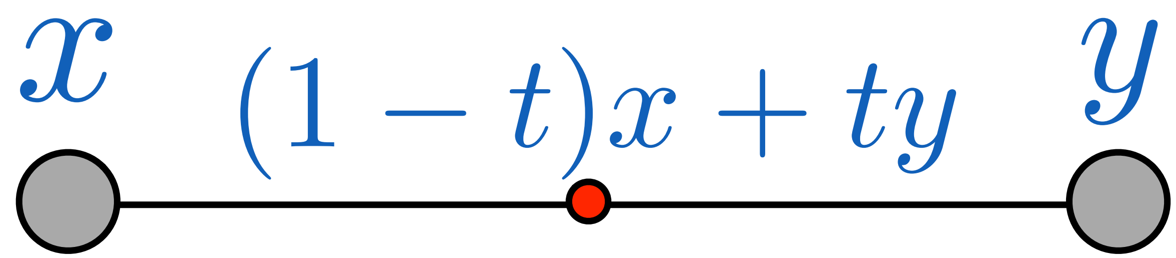


see also: [Rápcsák 1984; Udriste 1994]

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

The idea of geodesic convexity

Convexity



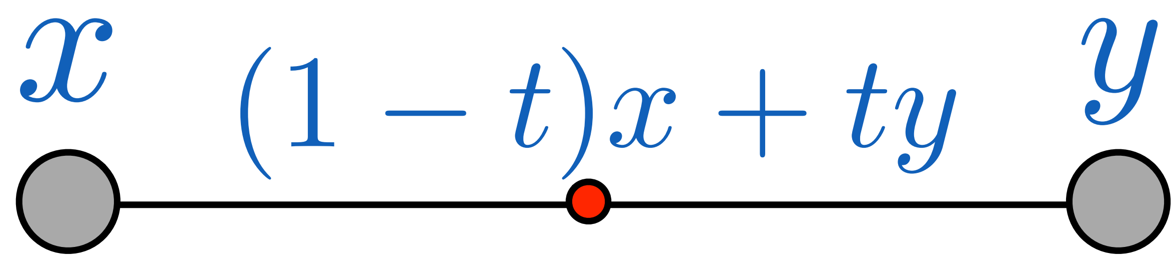
$$f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

see also: [Rápcsák 1984; Udriste 1994]

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

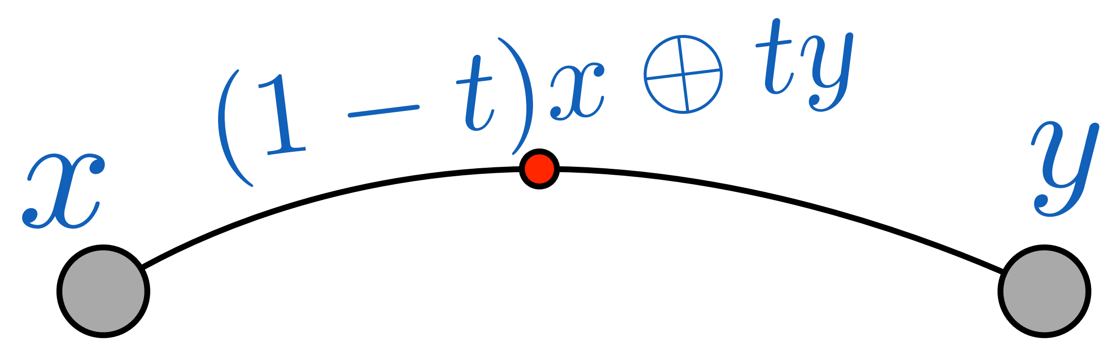
The idea of geodesic convexity

Convexity



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

Geodesic convexity

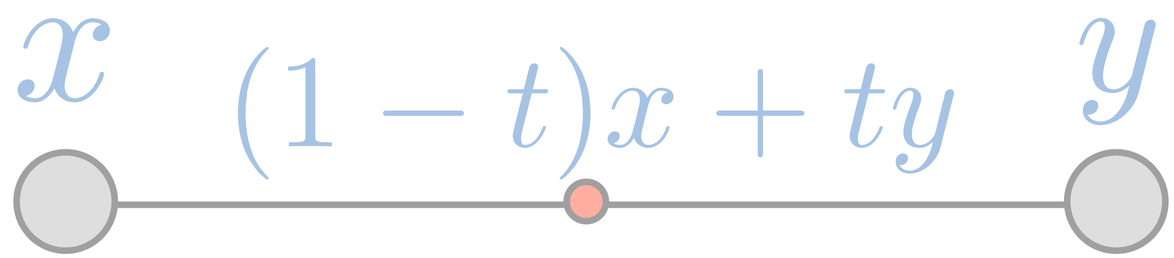


see also: [Rápcsák 1984; Udriste 1994]

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

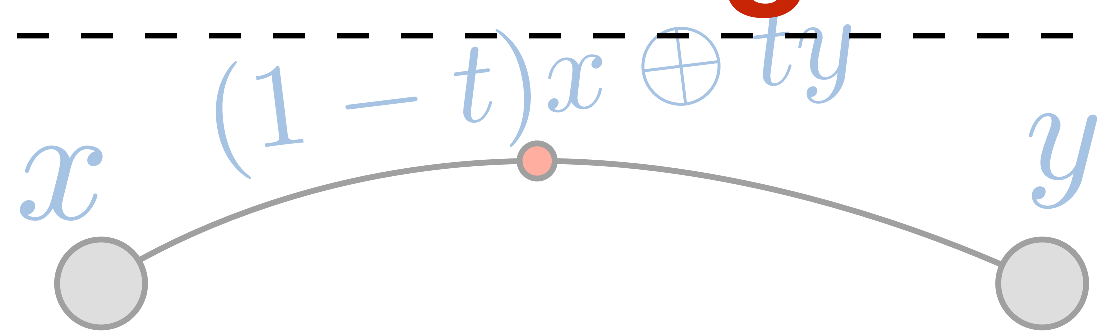
The idea of geodesic convexity

Convexity



Local opt of g-convex is global opt

Geodesic convexity

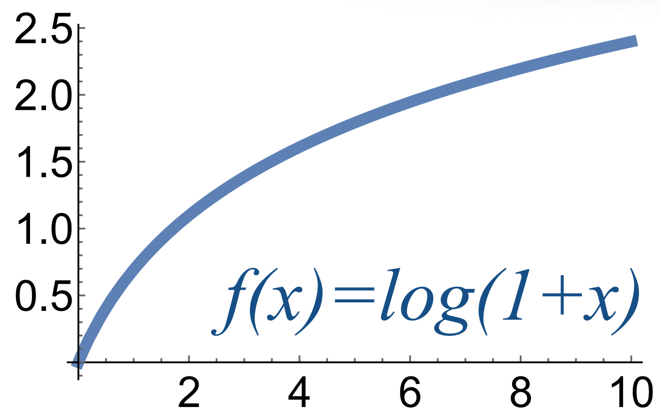


$$f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

see also: [Rápcsák 1984; Udriste 1994]

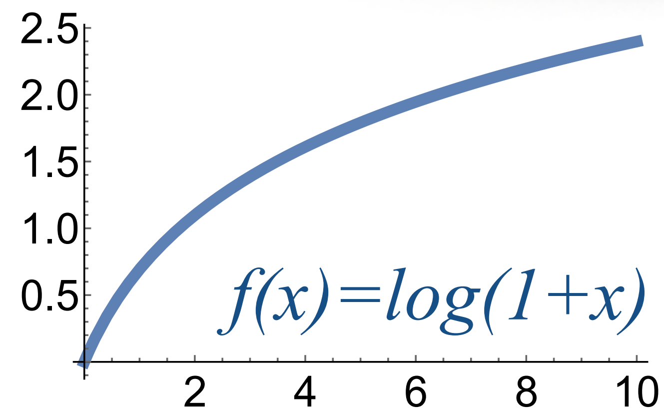
Metric spaces & curvature: [Menger; Alexandrov; Busemann; Bridson, Haefliger; Gromov; Perelman]

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

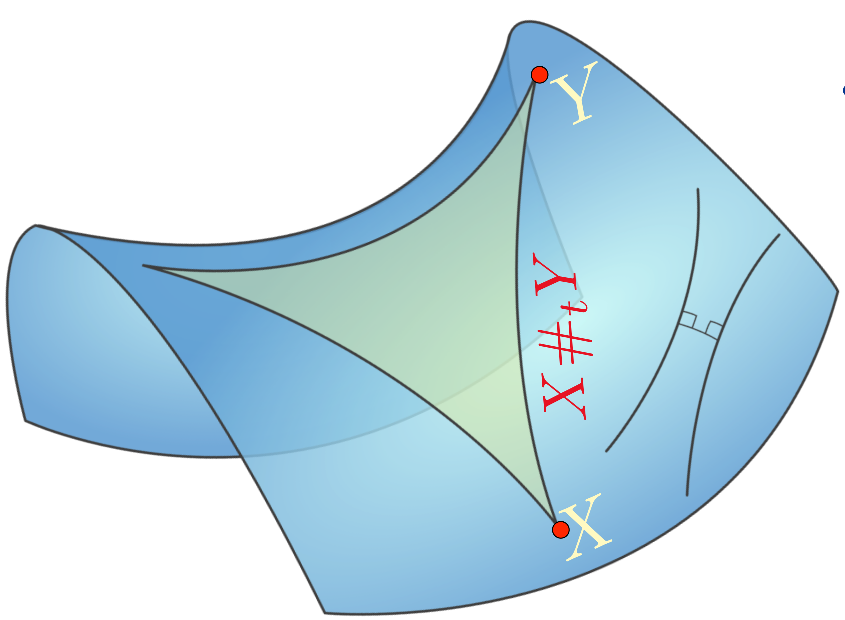


Example: *log(1+x)* concave in the usual sense, but geodesically convex since $f(x^{1-t}y^t) \leq (1-t)f(x)+tf(y)$

Geodesic from *X* to *Y*

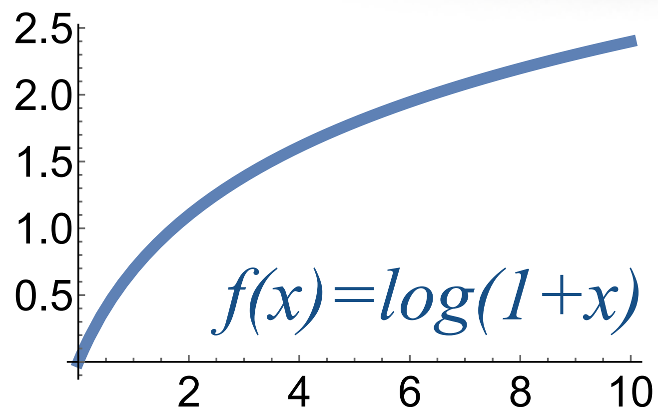
$$\gamma(t) \equiv (1-t)X \oplus tY := X^{\frac{1}{2}}(X^{-\frac{1}{2}}YX^{-\frac{1}{2}})^t X^{\frac{1}{2}}$$

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



Since $XY \neq YX$, cannot simply use $X^{1-t}Y^t$ as for scalars

G-convexity for positive definite matrices

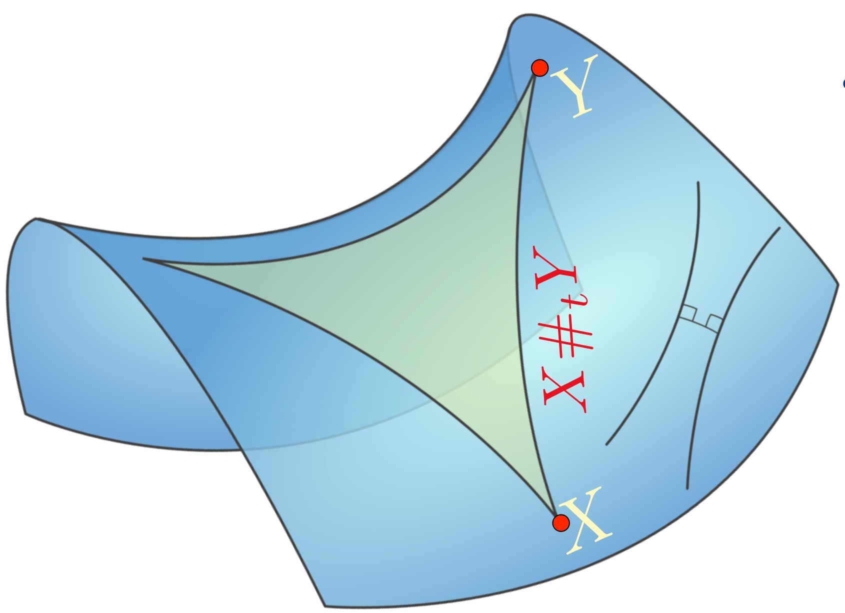


Example: *log(1+x)* concave in the usual sense, but geodesically convex since $f(x^{1-t}y^t) \leq (1-t)f(x)+tf(y)$

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

Important examples

Entropy of Gaussian, negative of log-barrier

$$f(X) = \log\det(X)$$

Euclidean concave but g-convex!

Condition number

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

Euclidean quasiconvex but g-convex

Generalized eigenvalue!

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

Euclidean quasiconvex
[Boyd, Ghaoui 1993; Nesterov, Nemirovksi 1991]

[Sra, Hosseini 2015; Sra 2017]

Important examples

Entropy of Gaussian, negative of log-barrier

$$f(X) = \log\det(X)$$

Euclidean concave but g-convex!

Condition number

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

Euclidean quasiconvex but g-convex

Generalized eigenvalue!

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

Euclidean quasiconvex
[Boyd, Ghaoui 1993; Nesterov, Nemirovksi 1991]

Many more!

[Sra, Hosseini 2015; Sra 2017]

G-convexity for positive def. matrices

Recognizing, constructing, and optimizing g-convex functions for positive def.

[Sra, Hosseini (2013,2015)]
[Sra 2017]

Several useful tools in there!

Corollaries

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

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

$$(X, Y) \mapsto \lambda_{\max}(XY)$$

Many more theorems and corollaries

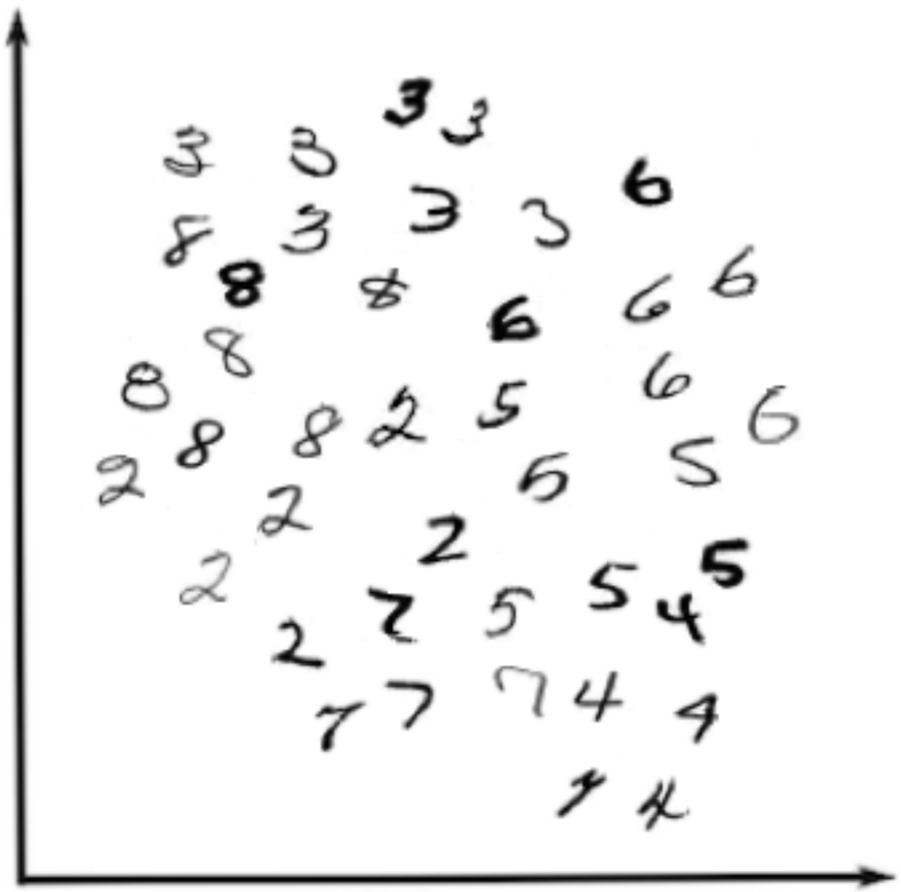
One-D version: **Geometric Programming**
www.stanford.edu/~boyd/papers/gp_tutorial.html
[Boyd, Kim, Vandenberghe, Hassibi (2007). 61pp.]

Geometry in Action

- Geodesically convex examples
- Non-geodesically convex examples

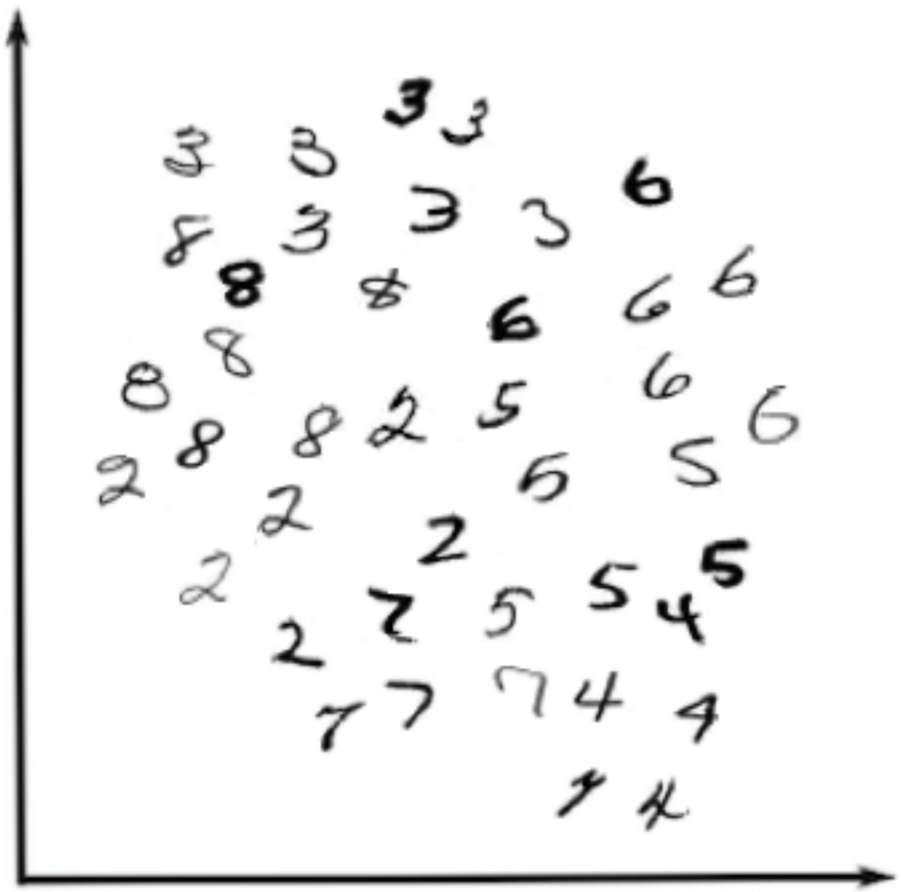
A new look at metric learning

Metric learning: a fundamental problem in machine learning



A new look at metric learning

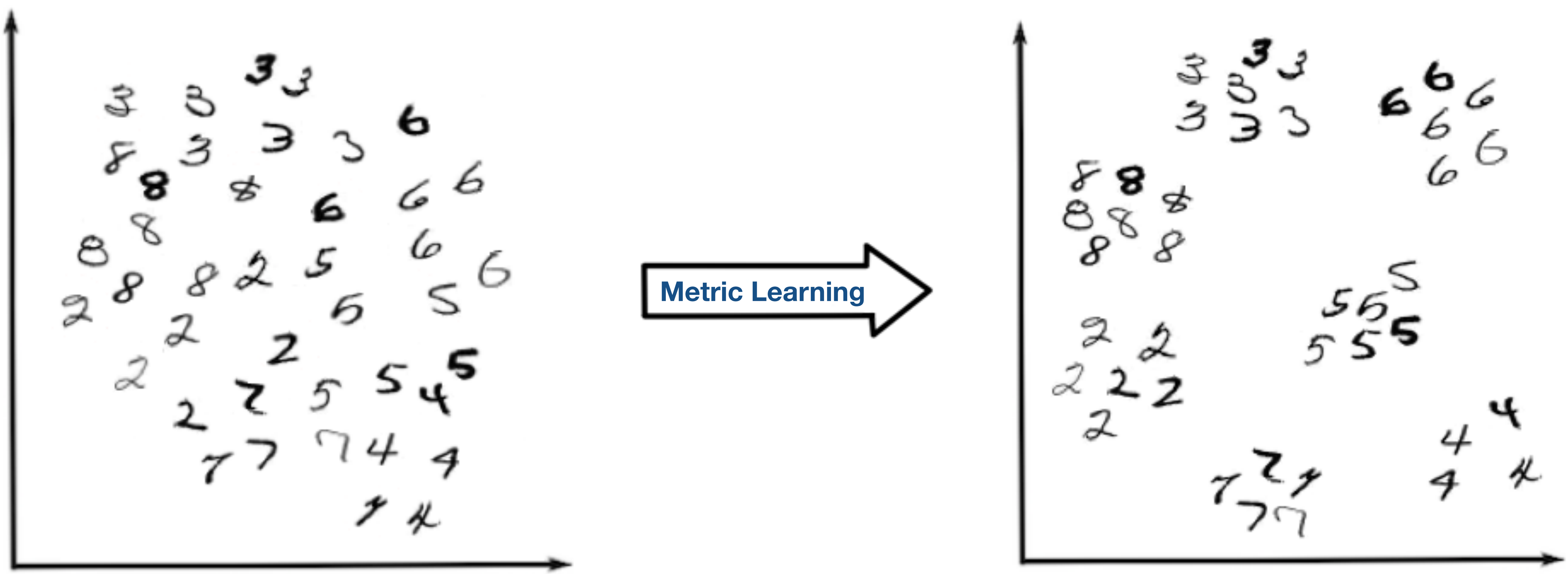
Metric learning: a fundamental problem in machine learning



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

A new look at 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)

A new look at 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

A new look at 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

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)

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

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

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

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)

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

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

LMNN

[Weinberger, Saul 2005]

large-margin SDP

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

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

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

ITML

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

relative entropy b/w Gaussians

$$\min_{\boldsymbol{A} \succeq 0} \quad D_{\mathrm{ld}}(\boldsymbol{A}, \boldsymbol{A}_0)$$

$$\text{such that} \quad d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) \leq u, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S},$$

$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) \geq l, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$$

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

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

Metric learning - convex formulations

MMC

[Xing, Jordan, Russell, Ng 2002]

Semidef. Programming (SDP)

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

LMNN

[Weinberger, Saul 2005]

large-margin SDP

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

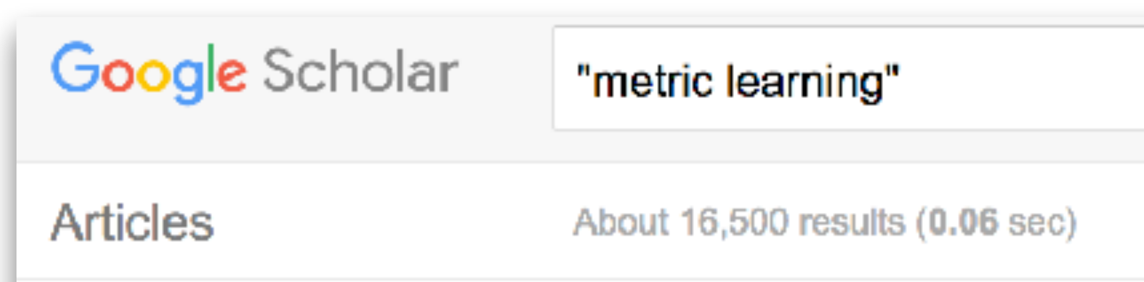
ITML

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

relative entropy b/w Gaussians

$$\min_{\boldsymbol{A} \succeq 0} D_{\mathrm{ld}}(\boldsymbol{A}, \boldsymbol{A}_0)$$
$$\text{such that} \quad d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) \leq u, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{S},$$
$$d_{\boldsymbol{A}}(\boldsymbol{x}, \boldsymbol{y}) \geq l, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{D}$$
$$D_{\mathrm{ld}}(\boldsymbol{A}, \boldsymbol{A}_0) := \mathrm{tr}(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - \log\det(\boldsymbol{A}\boldsymbol{A}_0^{-1}) - d$$

Tons of other works



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

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

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

A new geometric approach

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

Euclidean idea

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

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

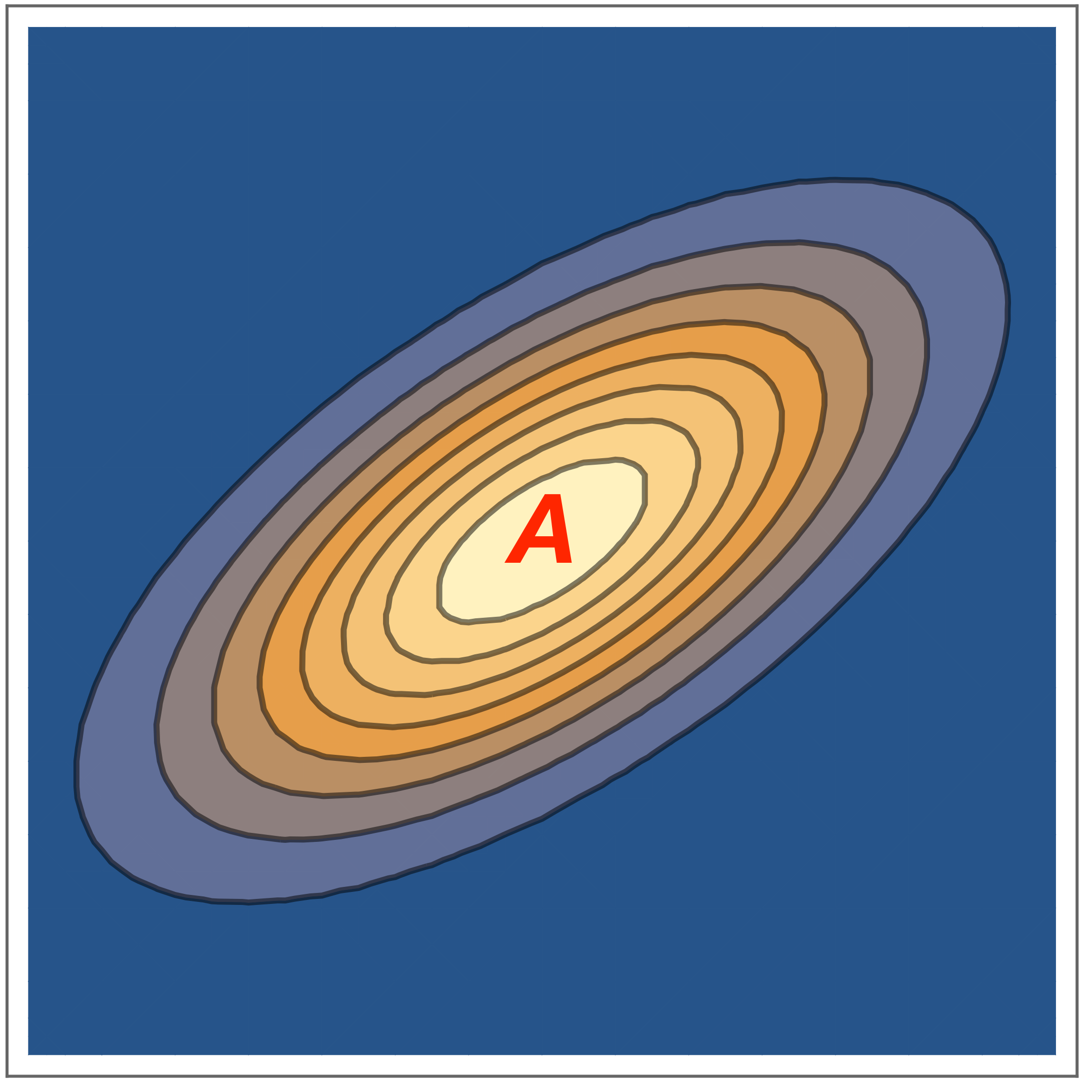
$$\min_{\boldsymbol{A} \succeq 0} \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{S}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j) - \lambda \sum_{(\boldsymbol{x}_i, \boldsymbol{x}_j) \in \mathcal{D}} d_{\boldsymbol{A}}(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

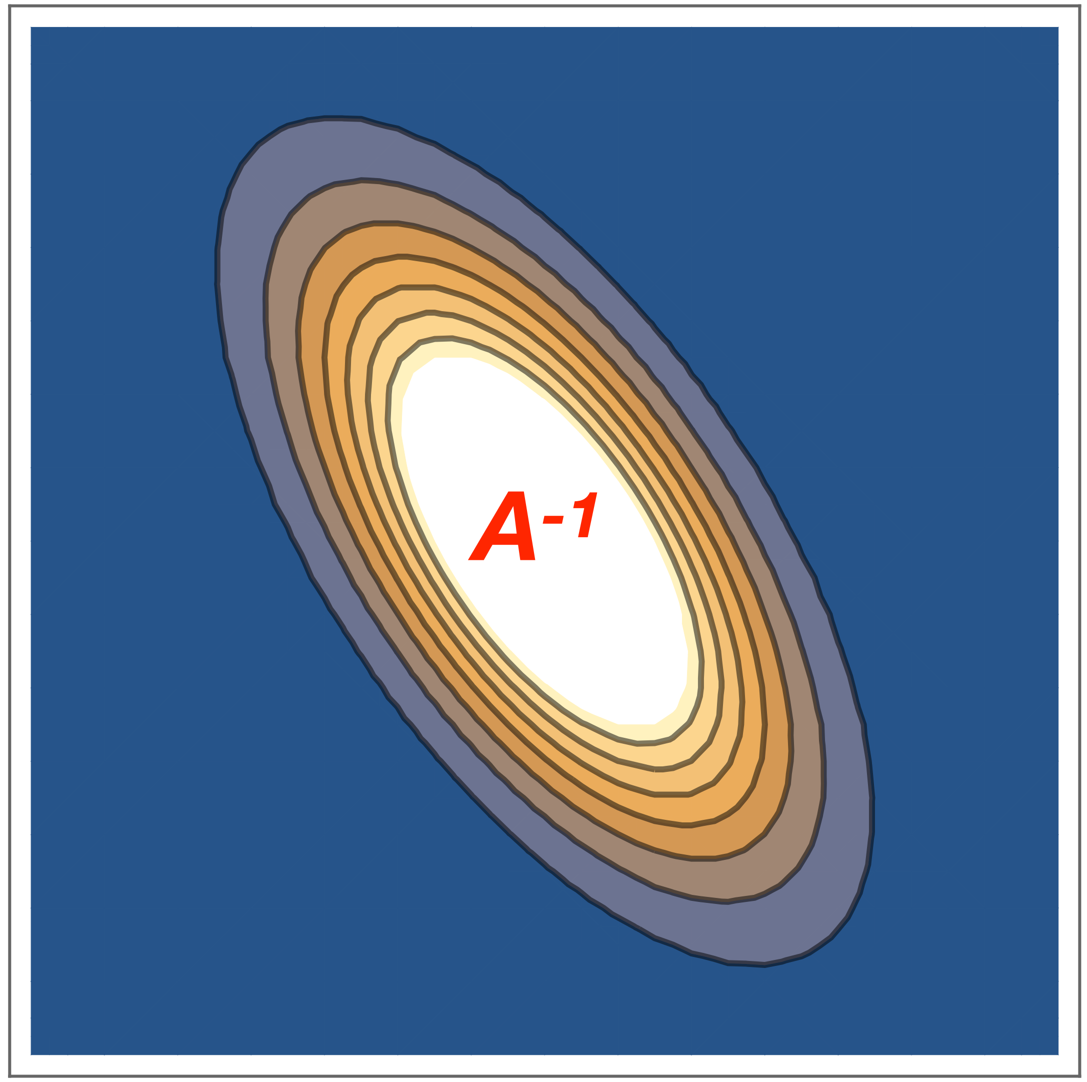
New idea

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

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

A new geometric approach





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

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

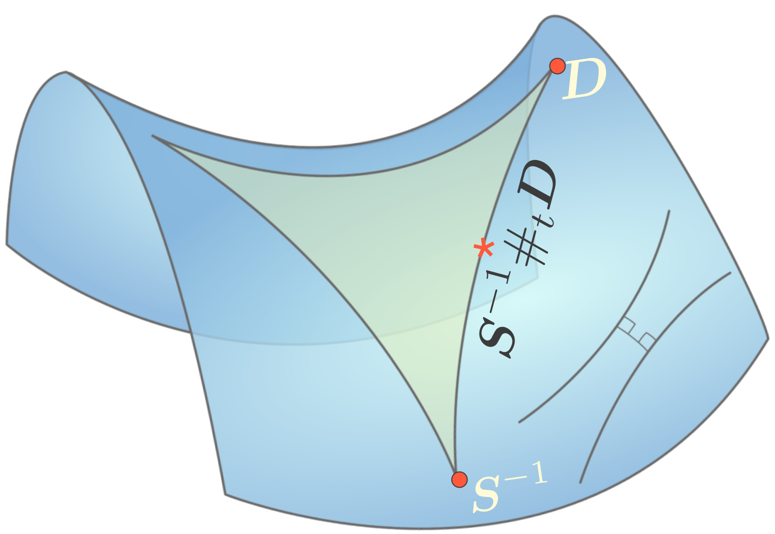
Closed form solution!

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

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

More generally

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



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

Geometric approach to metric learning

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

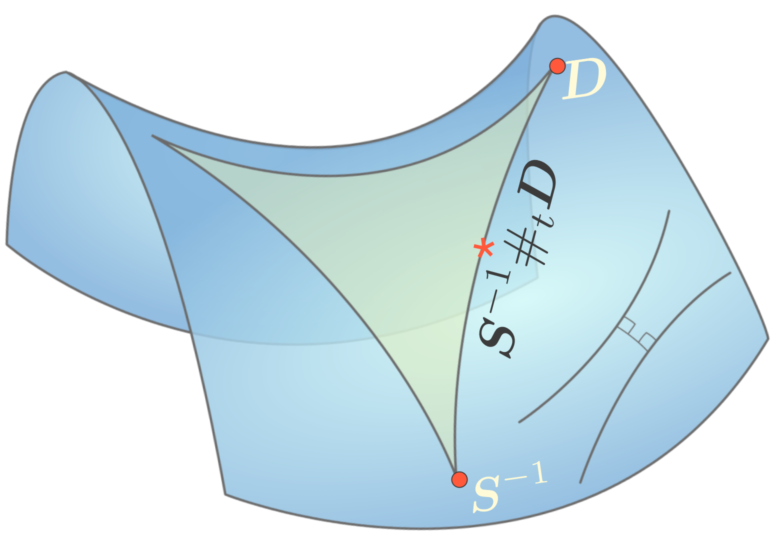
Closed form solution!

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

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

More generally

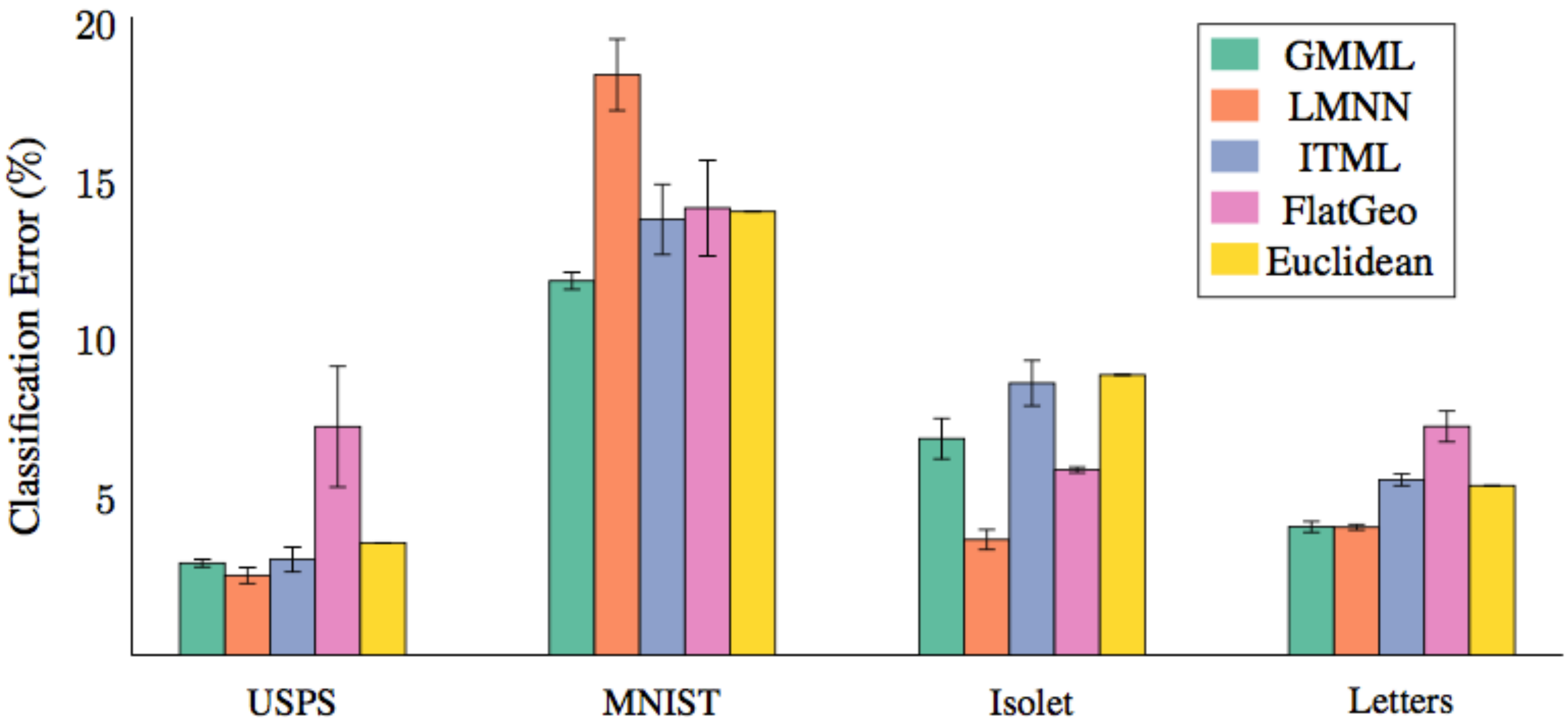
$$\min_{\boldsymbol{A}\succ 0} \quad (1-t)\delta_R^2(\boldsymbol{S}^{-1}, \boldsymbol{A}) + t\delta_R^2(\boldsymbol{D}, \boldsymbol{A})$$



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

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

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

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

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

Brascamp-Lieb Constant

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

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

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

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

Brascamp-Lieb constant

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

- 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

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

- Applications to geometric complexity theory

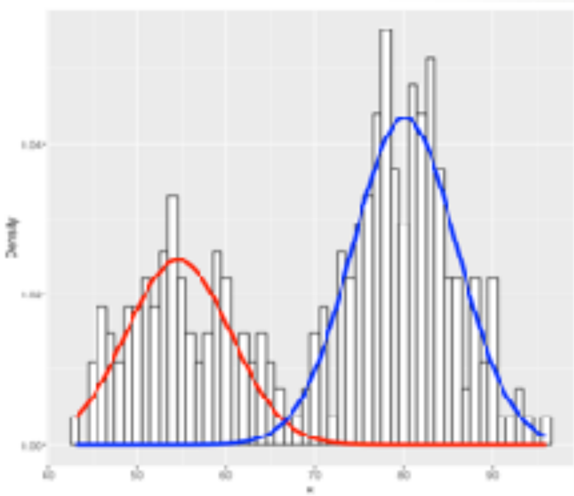
  [Garg, Gurvits, Oliveira, Wigderson; Jul 2016]
- Problem has unique solution & sufficient conditions

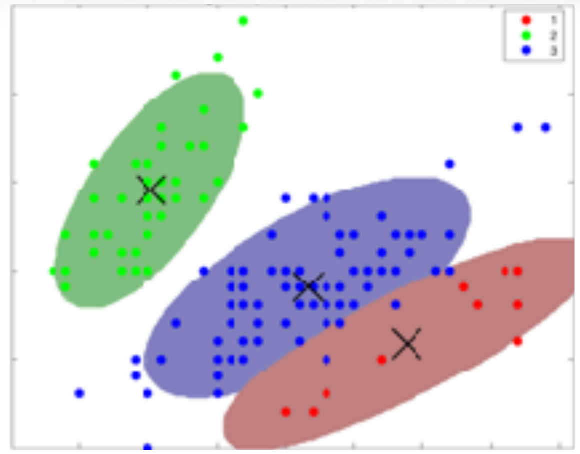
  [Bennett, Carbery, Christ, Tao, 2005]
- Barthe, Carlen, Lieb, Cordero-Erasquin, McCann, …

Prop: This is a g-convex optimization problem

Proof: Corollary 2.11 in *[Sra, Hosseini, 2015]*

Gaussian mixture models

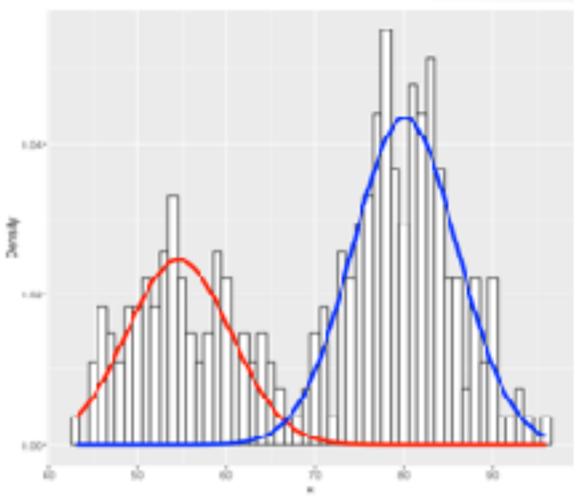


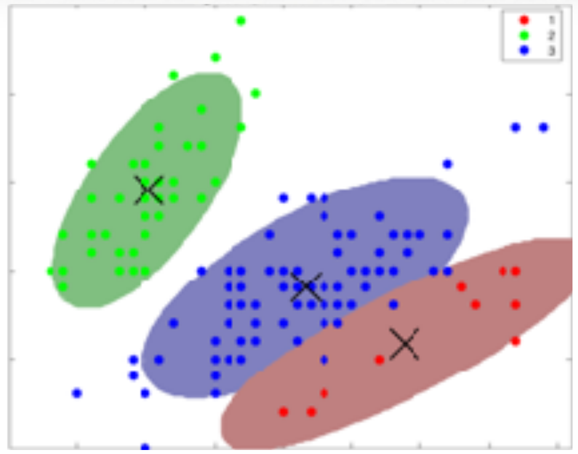


$$p(x) = \sum_k \pi_k \text{Gaussian}(x; \mu_k, \Sigma_k)$$

Aim: Given training data $x_1, \ldots, x_n$, estimate μ_k, Σ_k

Gaussian mixture models

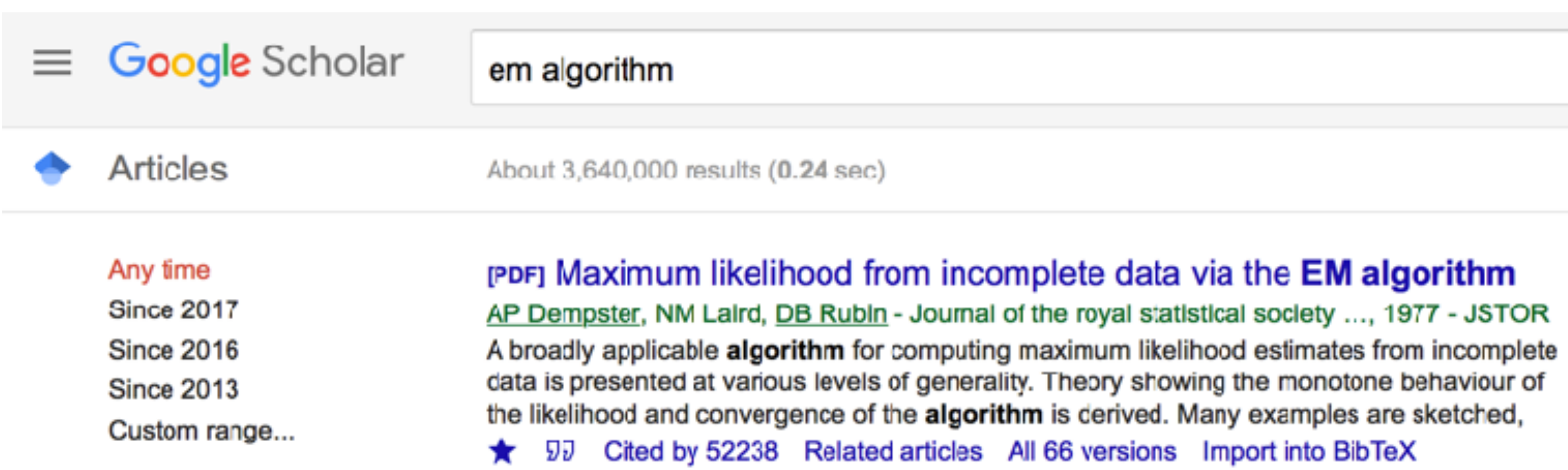




$$p(x) = \sum_k \pi_k \text{Gaussian}(x; \mu_k, \Sigma_k)$$

Aim: Given training data $x_1, \ldots, x_n$, estimate μ_k, Σ_k

Expectation maximization (EM): default choice



Gaussian mixture models

- **Nonconvex –** difficult, possibly several local optima
- **Theory -** Recent progress (Moitra, Valiant 2010; Daskalakis et al, 2017; more!)
- **In Practice –** EM still default choice, for it posdef is easy

Other methods: How to incorporate the positive definiteness constraint on Σ_k

[Hosseini, Sra NIPS 2015]

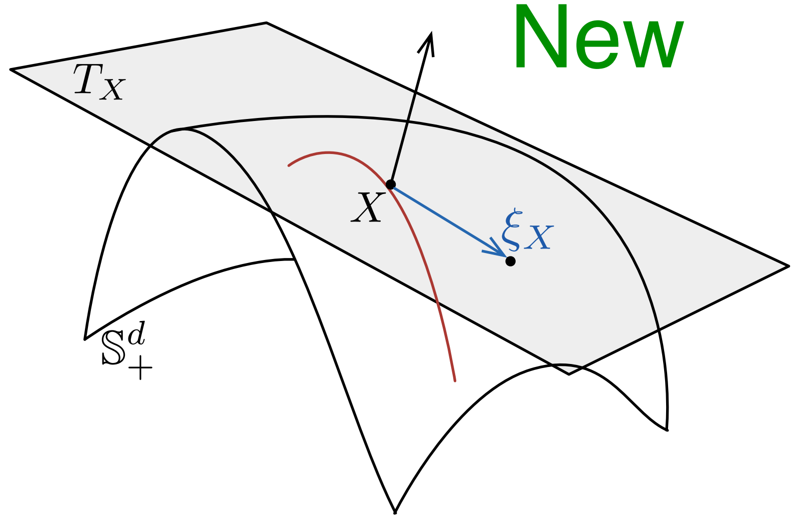
Gaussian mixture models

- **Nonconvex –** difficult, possibly several local optima
- **Theory -** Recent progress (Moitra, Valiant 2010; Daskalakis et al, 2017; more!)
- **In Practice –** EM still default choice, for it posdef is easy

Other methods: How to incorporate the positive definiteness constraint on Σ_k

Geometric opt.

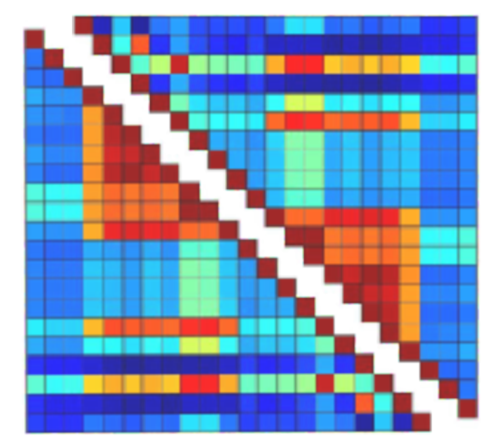
New



Unconstrained, Cholesky

Folklore

LL^T



[Hosseini, Sra NIPS 2015]

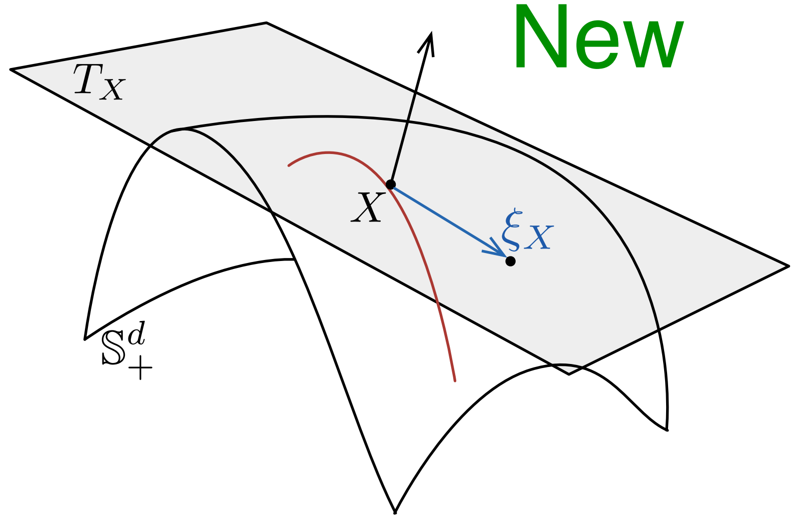
Gaussian mixture models

- **Nonconvex –** difficult, possibly several local optima
- **Theory -** Recent progress (Moitra, Valiant 2010; Daskalakis et al, 2017; more!)
- **In Practice –** EM still default choice, for it posdef is easy

Other methods: How to incorporate the positive definiteness constraint on Σ_k

Geometric opt.

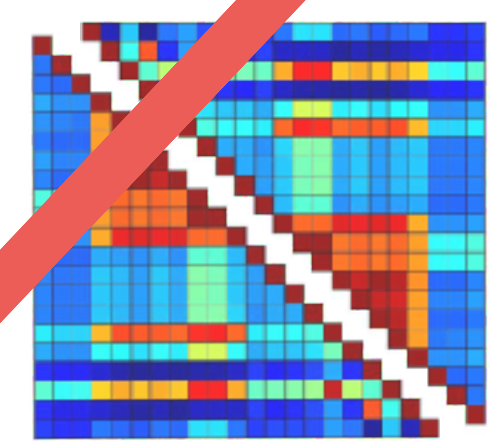
New



Unconstrained, Cholesky

Folklore

$$LL^T$$



[Hosseini, Sra NIPS 2015]

Naive use of Riemannian opt. fails!

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

Showing "time // negative log-likelihood (avg)"

manopt.org

Riemannian opt. toolbox

d=35

n=200,000

A better formulation?

A better formulation?

log-likelihood for one component

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

A better formulation?

log-likelihood for one component

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

Euclidean convex problem
Not geodesically convex

A better formulation?

log-likelihood for one component

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

Euclidean convex problem
Not geodesically convex

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.

Reaping the benefits of geometry

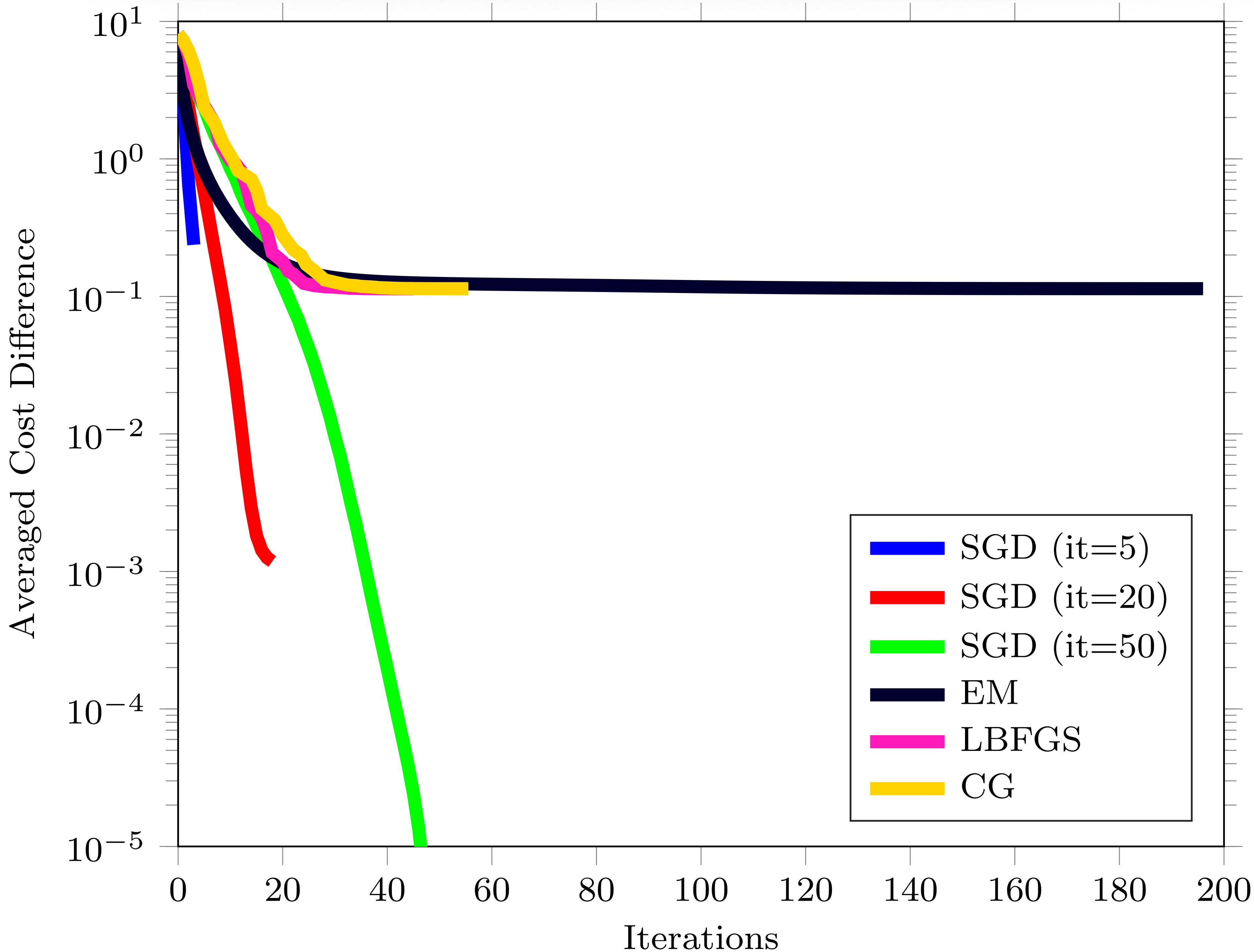
K	EM	Our manifold LBFGS
2	17s // 29.28	**14s** // 29.28
5	202s // 32.07	**117s** // 32.07
10	2159s // 33.05	**658s** // 33.06

Showing "time // negative log-likelihood (avg)"

d=35
n=200,000

github.com/utvisionlab/mixest

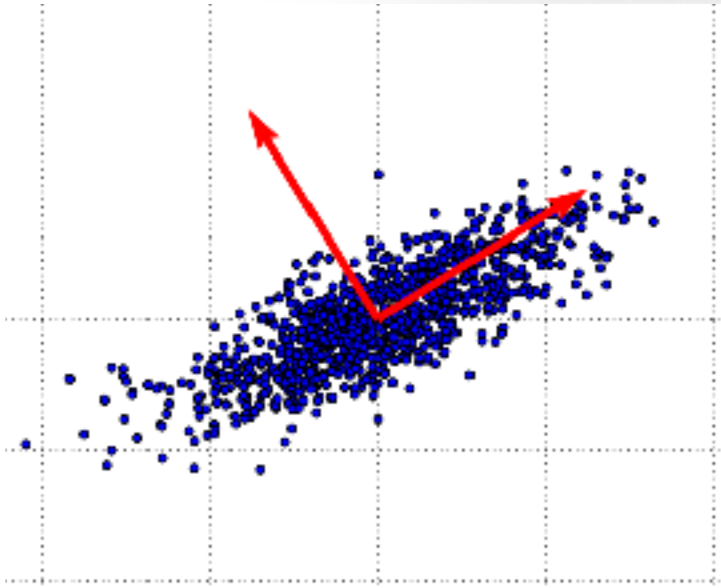
Large-scale: use Riemannian SGD



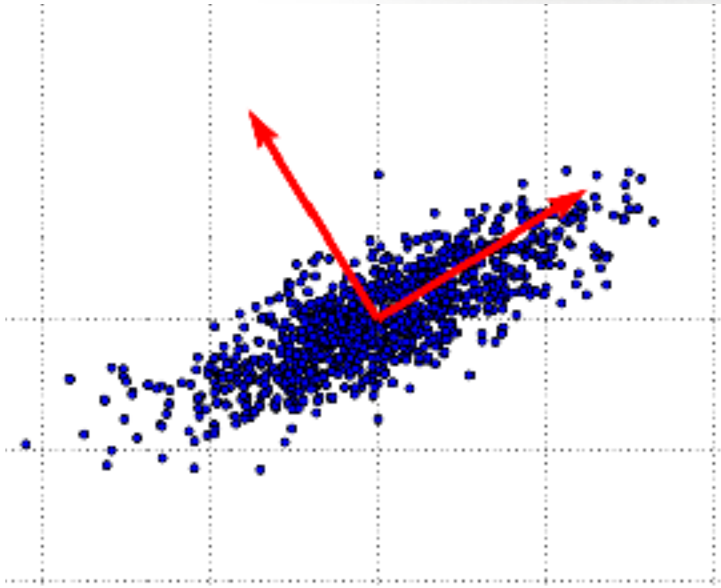
(d=90, n=515345, k=7)

[Hosseini, Sra, 2017]

PCA for large datasets



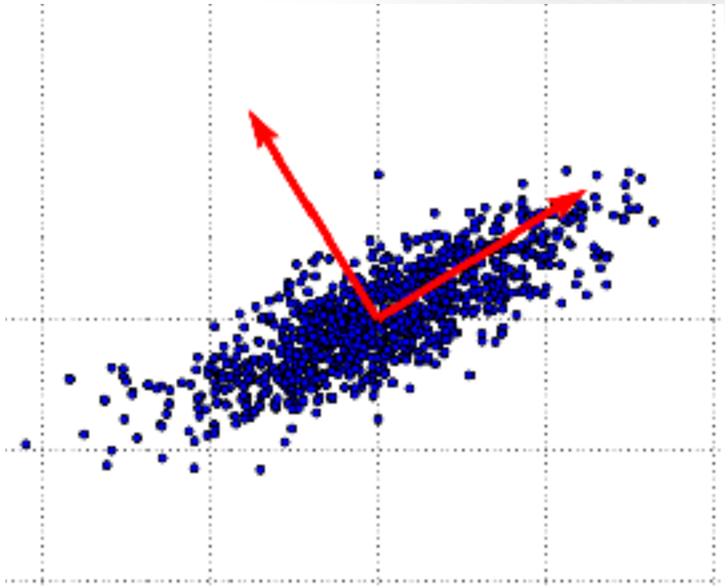
PCA for large datasets



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

n is big

PCA for large datasets



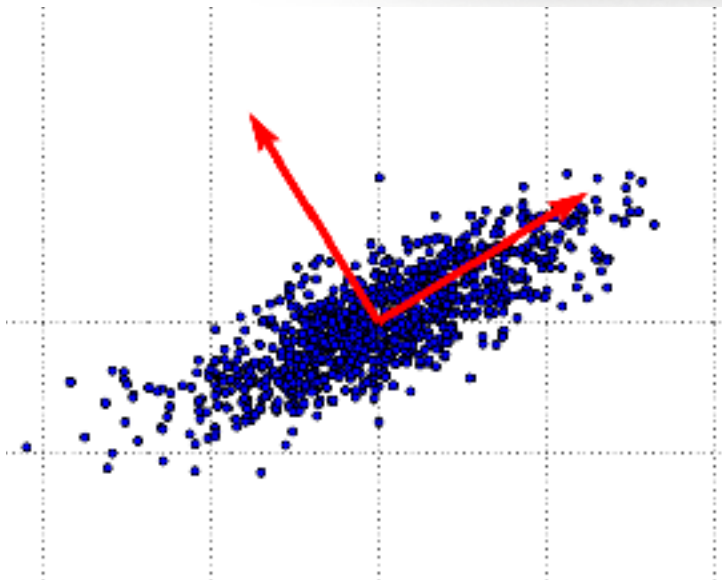
$$\min_{x^T x = 1} \quad -x^T \left(\sum_{i=1}^{n} z_i z_i^T\right) x$$

n is big

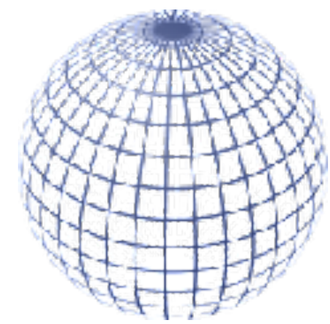
Lots of recent work on "SGD" for eigenvectors

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

PCA for large datasets



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



n is big

Lots of recent work on “SGD” for eigenvectors

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

Simpler analysis thanks to a key geometric realization

Even though the problem is geodesically non-convex
it “behaves like” geodesically convex on the sphere.

→ Running Riemannian SGD will obtain global optimum

[Zhang, Reddi, Sra, NIPS 2016]

Summary: geometry in action

1. Simple geometric model for metric learning

   (vastly faster, cleaner than traditional formulations!)
2. Geometry guided reformulation + algo for GMMs
3. Insights into why we can solve large-scale PCA

All three are nonconvex; geodesic convexity plays a crucial role

Theory

Theory

Theory for first-order optimization

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

Theory for first-order optimization

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

Assume: we can obtain exact or stochastic gradients

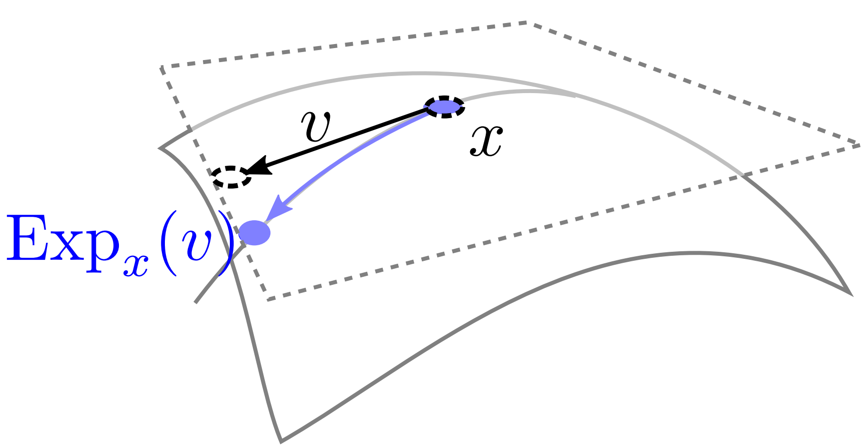
Theory for first-order optimization

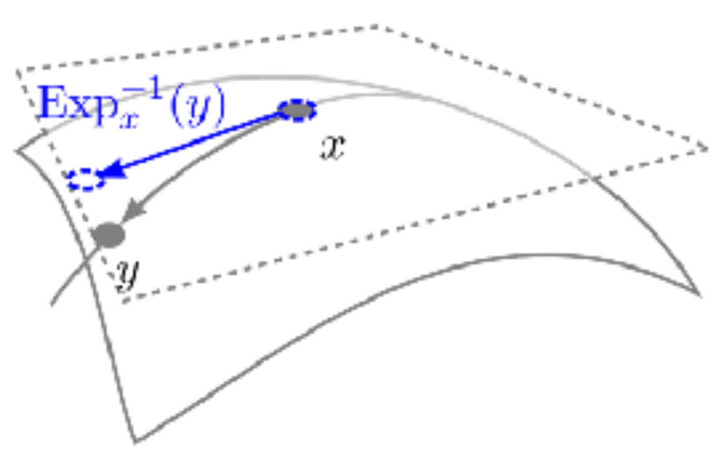
$$\min_{x \in \mathcal{X} \subset \mathcal{M}} \quad f(x)$$

Assume: we can obtain exact or stochastic gradients

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

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





Aim: Develop global complexity theory of first-order g-convex optimization

Aim: Develop global complexity theory of first-order g-convex optimization

Global Complexity

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

Convex Optimization

[Nemirovski-Yudin 1983]
[Nesterov 2003]
Le Roux, Schmidt, Bach; Gurbuzbalaban, Ozdaglar, Parrilo; Defazio et al;

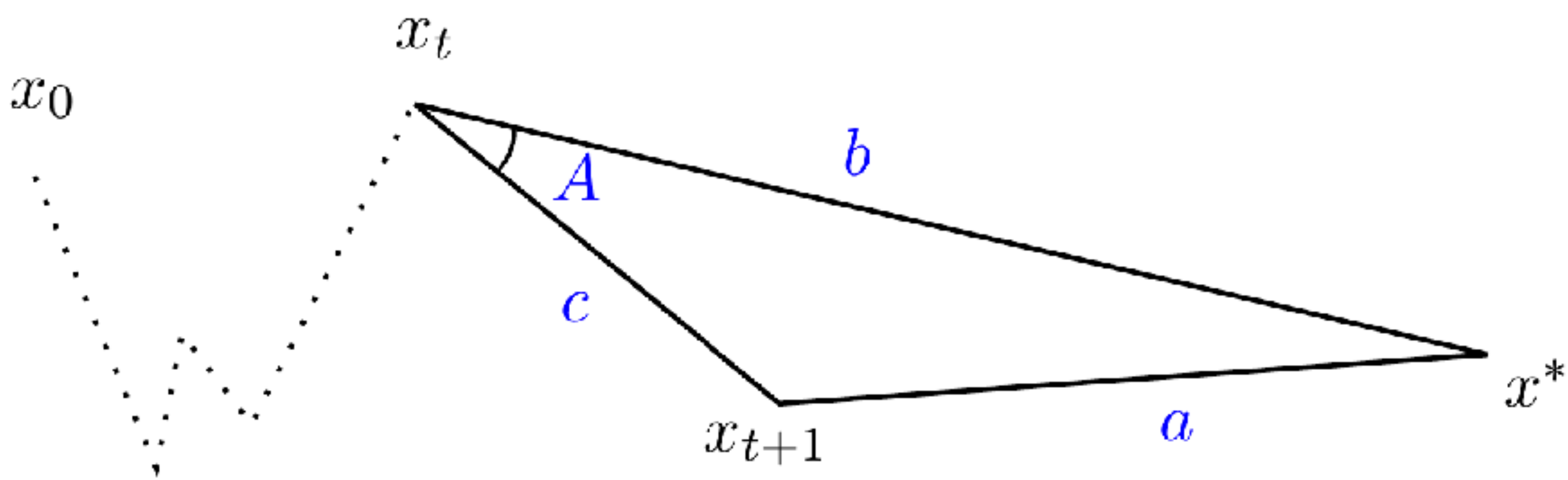
G-Convex Optimization

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

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

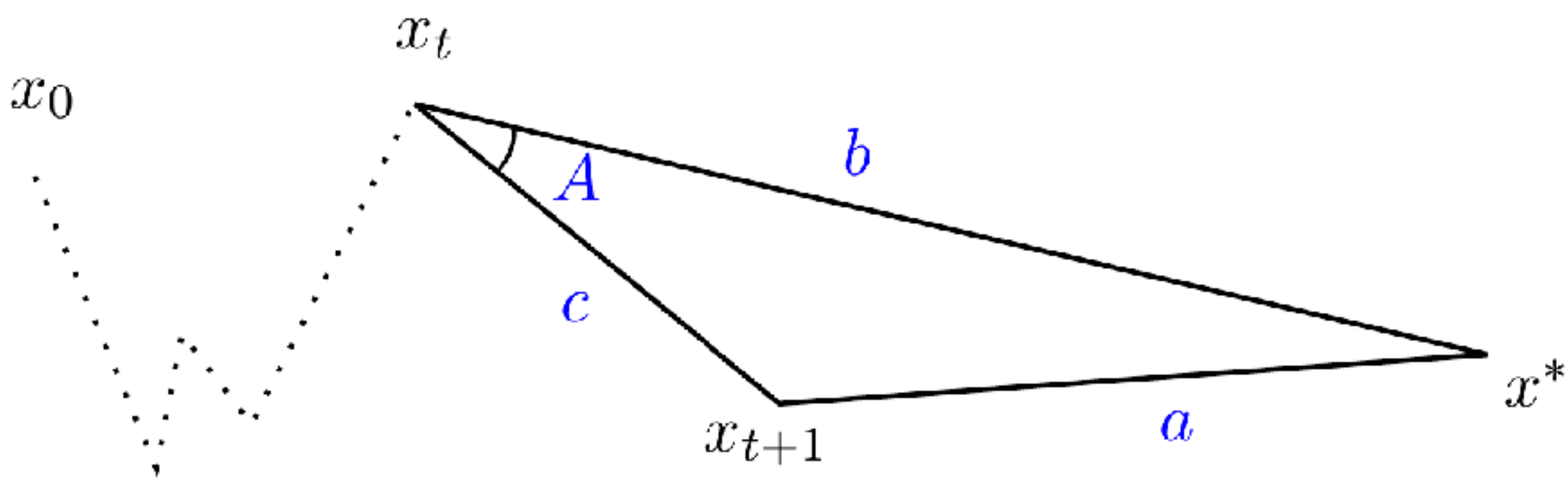
The Euclidean law of cosines is essential to bound $d^2(x_{t+1}, x^*)$ in analysis of usual convex opt. methods

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



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

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



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

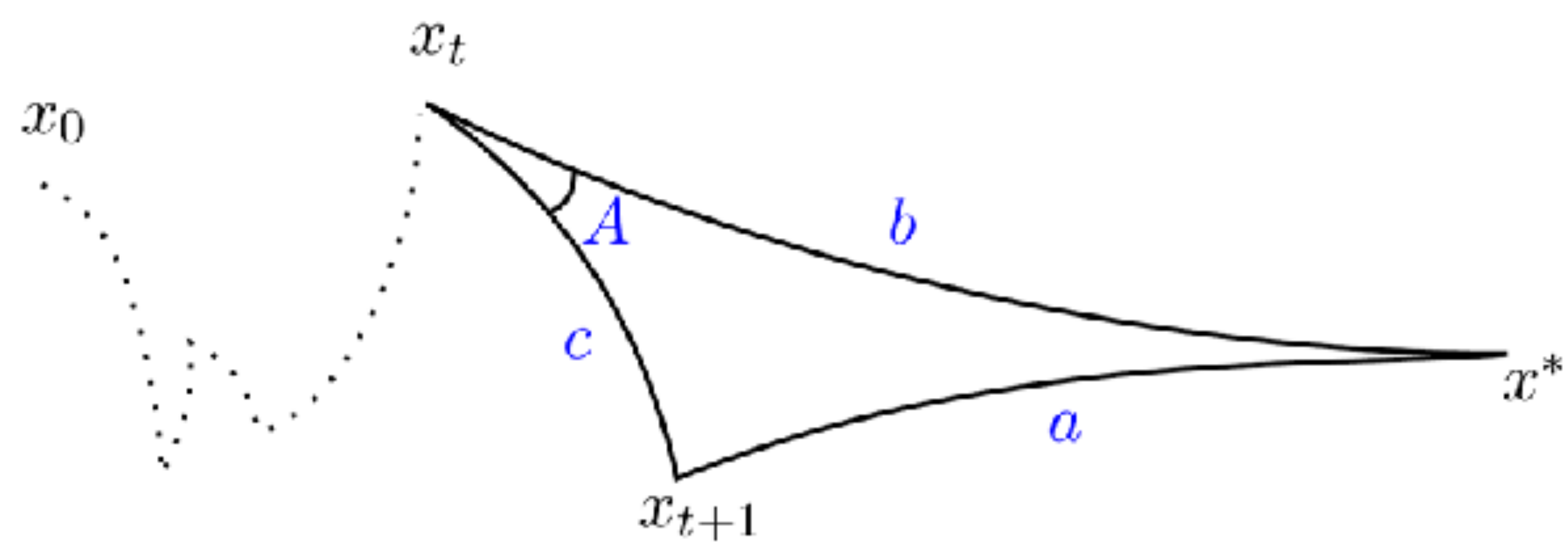
$$\Downarrow$$

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

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

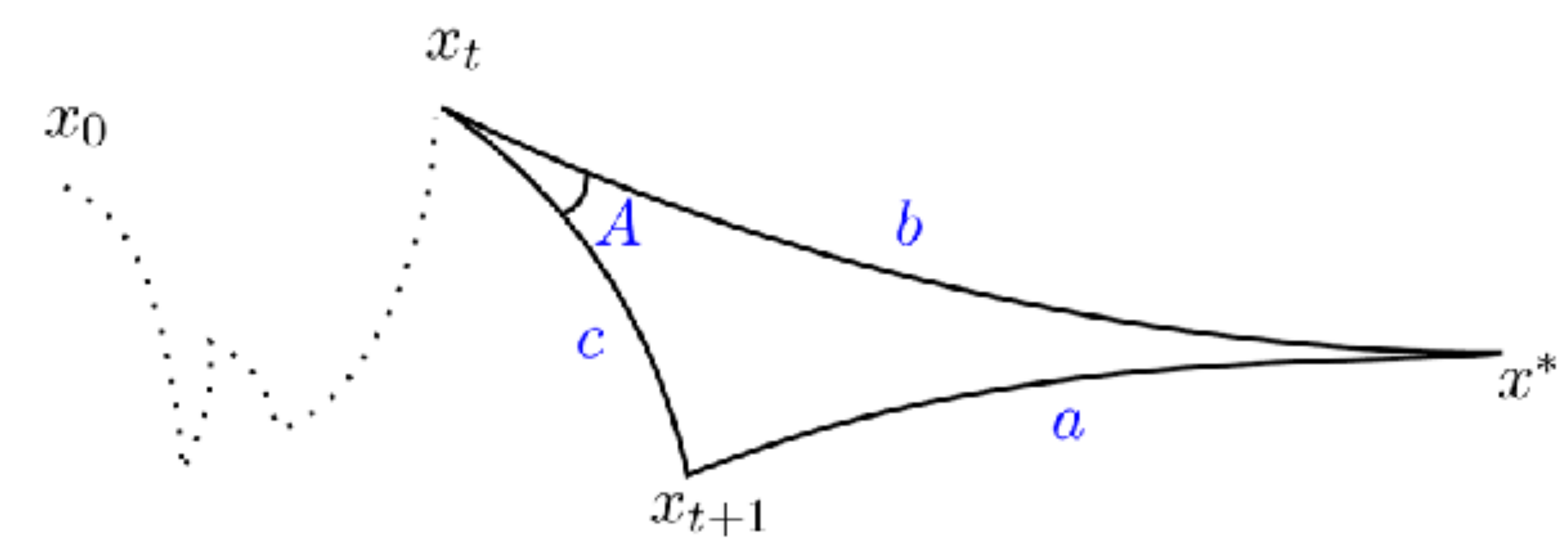
[Zhang, Sra, COLT 2016]

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



[Zhang, Sra, COLT 2016]

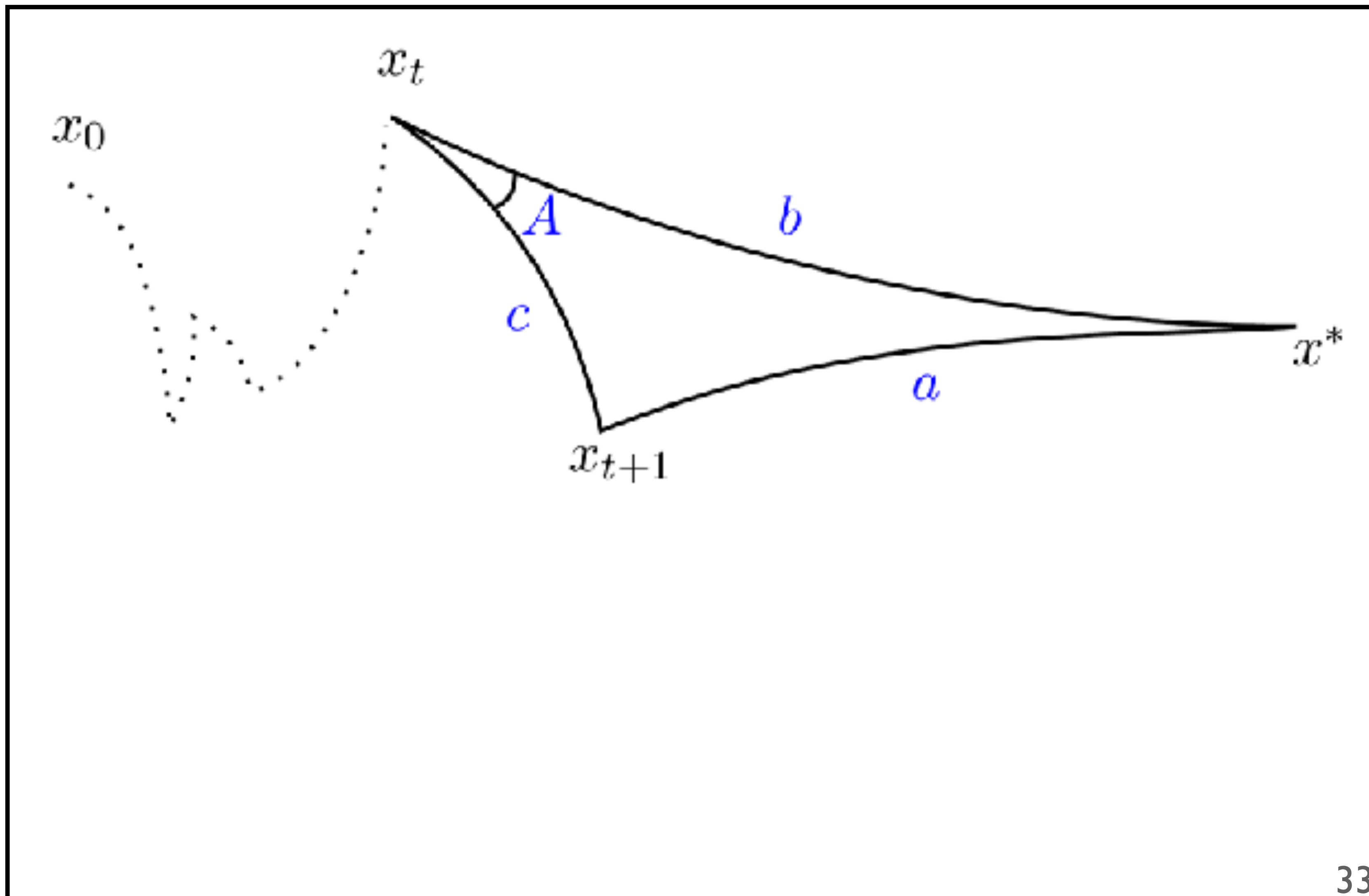
We develop a corresponding inequality to bound $d^2(x_{t+1}, x^*)$ on manifolds (and related spaces)



[Zhang, Sra, COLT 2016]

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

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

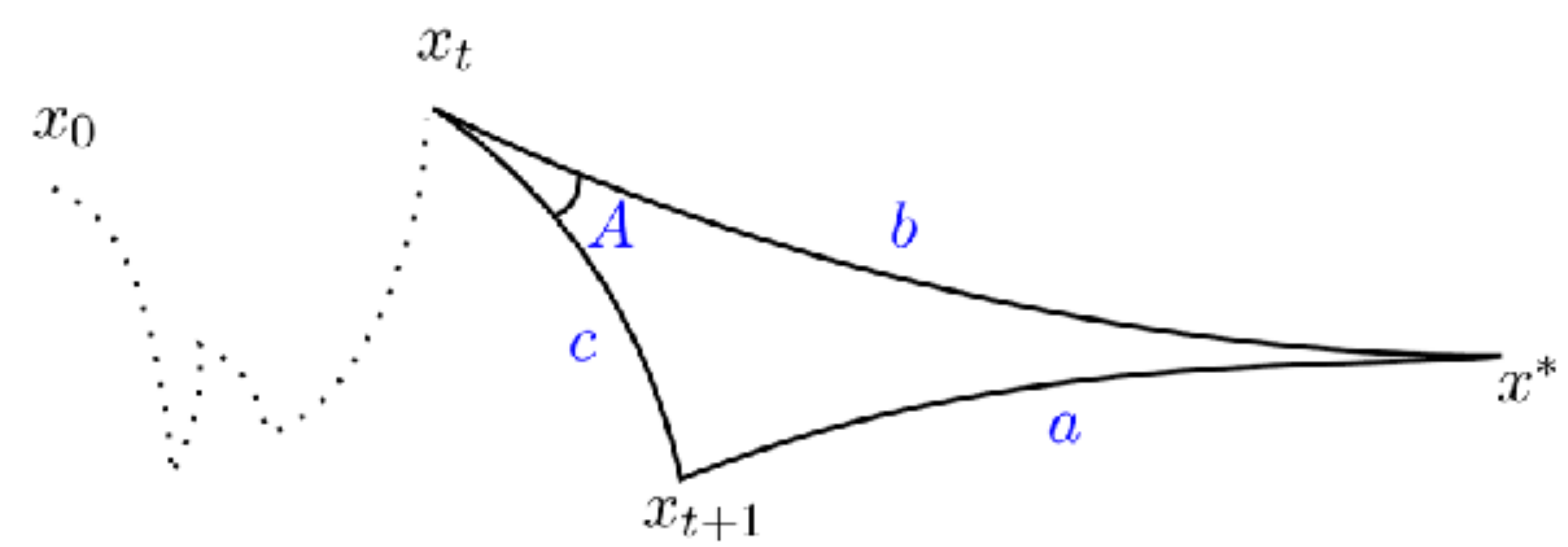


[Zhang, Sra, COLT 2016]

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

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

Grönwall's inequality



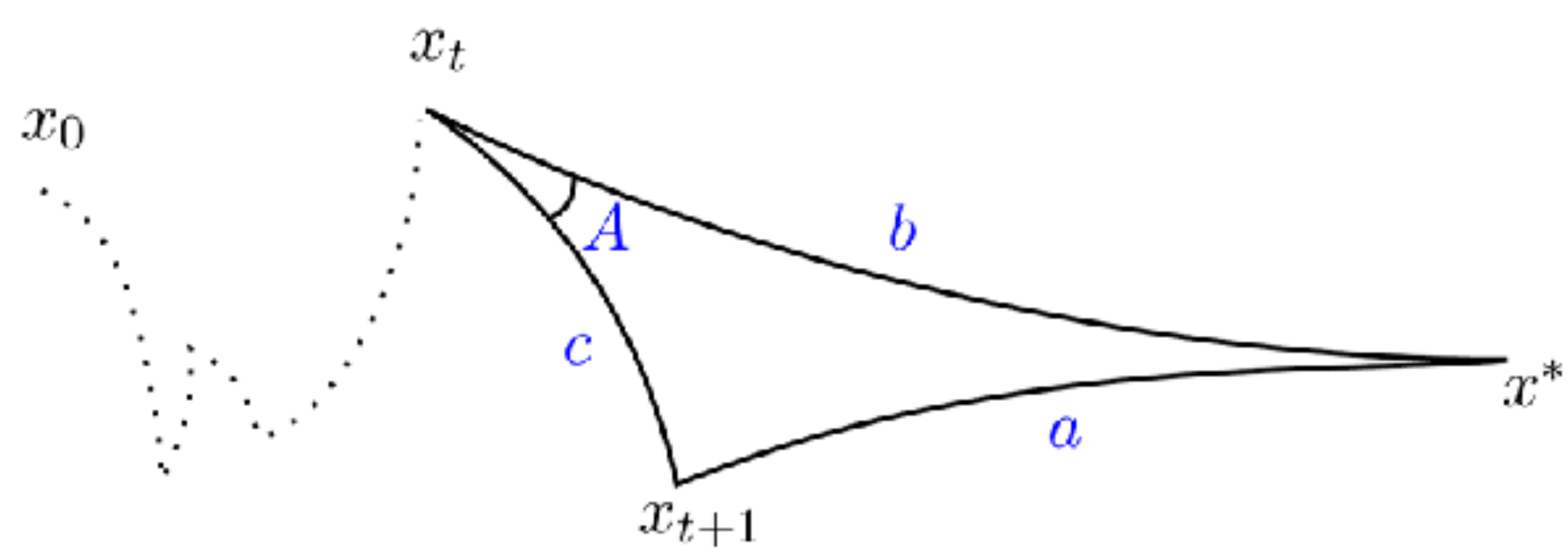
[Zhang, Sra, COLT 2016]

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

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

Toponogov's theorem

Grönwall's inequality



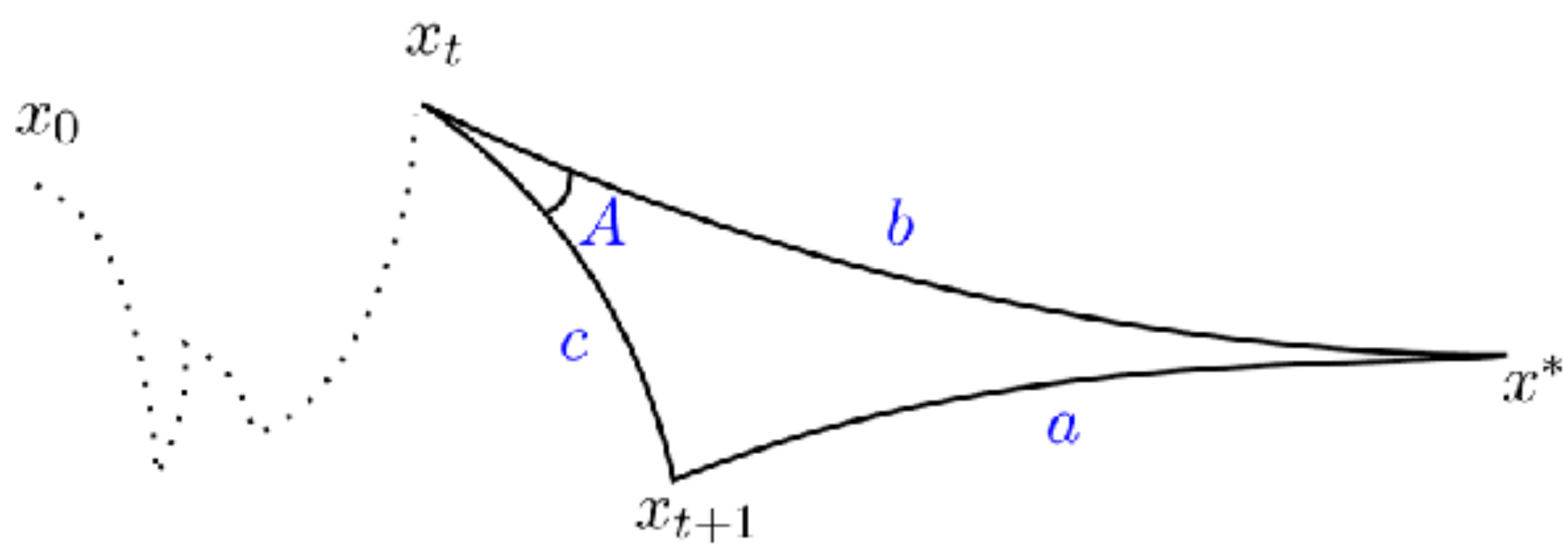
[Zhang, Sra, COLT 2016]

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

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

Toponogov's theorem

Grönwall's inequality



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

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

[Zhang, Sra, COLT 2016]

Rates depend on lower bounds on sectional curvature

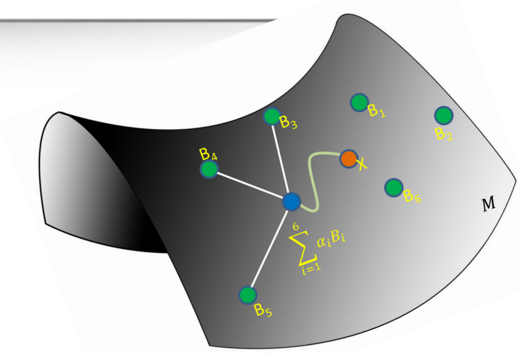
(Sub)gradient	convex	g-convex
Lipschitz	$O\left(\sqrt{\frac{1}{t}}\right)$	$O\left(\sqrt{\frac{\zeta_{\max}}{t}}\right)$
Strongly convex / smooth	$O\left(\frac{1}{t}\right)$	$O\left(\frac{\zeta_{\max}}{t}\right)$
Strongly convex & smooth	$O\left((1-\frac{\mu}{L_g})^t\right)$	$O\left((1-\min\{\frac{1}{\zeta_{\max}},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

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]*

Riemannian finite-sum problems



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

- $\mathcal{M}$ is a Riemannian manifold
- g-convex and g-nonconvex 'f' allowed
- First global complexity results for stochastic methods on Riemannian manifolds
- Riemannian SVRG

[Zhang, Reddi, Sra, NIPS 2016]

But some of the analysis does not trivially generalize...

Lemma: Let f be convex and L-smooth in a vector space, then

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof in textbook!

But some of the analysis does not trivially generalize...

Lemma: Let f be convex and L-smooth in a vector space, then

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof in textbook!

Lemma: Let f be g-convex and Riemannian-L-smooth, then

$$\|\mathrm{grad} f(x) - \Gamma_y^x \mathrm{grad} f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), \mathrm{Exp}_y^{-1}(x) \rangle)$$

But some of the analysis does not trivially generalize...

Lemma: Let f be convex and L-smooth in a vector space, then

$$\|\nabla f(x) - \nabla f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), x - y \rangle)$$

Proof in textbook!

Lemma: Let f be g-convex and Riemannian-L-smooth, then

$$\|\mathrm{grad} f(x) - \Gamma_y^x \mathrm{grad} f(y)\|^2 \leq 2L(f(x) - f(y) - \langle \nabla f(y), \mathrm{Exp}_y^{-1}(x) \rangle)$$

Proof broken!

Open problem

Summary of Riemannian SVRG results

f_i	$f = \sum_i f_i$	# Incremental First-order Oracle (IFO)
L-g-smooth	None	$O\left(n + n^{2/3}\zeta^{1/2}\frac{1}{\epsilon^2}\right)$
	τ-gradient dominated	$O((n + n^{2/3}\zeta^{1/2}\tau L)\log(\frac{1}{\epsilon}))$
	μ-strongly g-convex	$O\left((n + \frac{\zeta L^2}{\mu^2})\log(\frac{1}{\epsilon})\right)$ or $O\left((n + n^{2/3}\zeta^{1/2}\frac{L}{\mu})\log(\frac{1}{\epsilon})\right)$

Summary of Riemannian SVRG results

f_i	$f = \sum_i f_i$	# Incremental First-order Oracle (IFO)
L-g-smooth	None	$O\left(n + n^{2/3}\zeta^{1/2}\frac{1}{\epsilon^2}\right)$
	τ-gradient dominated	$O((n + n^{2/3}\zeta^{1/2}\tau L)\log(\frac{1}{\epsilon}))$
	μ-strongly g-convex	$O\left((n + \frac{\zeta L^2}{\mu^2})\log(\frac{1}{\epsilon})\right)$ or $O\left((n + n^{2/3}\zeta^{1/2}\frac{L}{\mu})\log(\frac{1}{\epsilon})\right)$

Same as SVRG and non-convex SVRG, except for ζ and worse constants in the g-convex case

Accelerated gradient on manifolds

An Estimate Sequence for Geodesically Convex Optimization.
Hongyi Zhang, Suvrit Sra.
31th Annual Conference on Learning Theory (COLT'18).

Riemannian Nesterov accelerates locally

First proof of acceleration on Riemannian manifolds

(informal) For μ-strongly g-convex, L-g-smooth functions, if the initialization is at most $\frac{1}{20\sqrt{K}}\left(\frac{\mu}{L}\right)^{\frac{3}{4}}$ away from x^*, then with properly chosen parameters, it takes $O\left(\sqrt{\frac{L}{\mu}}\log(\frac{1}{\epsilon})\right)$ gradient evaluations to reach ϵ accuracy.

Acceleration without strong g-convexity — Open problem

Global convergence of Riemannian Nesterov — Open problem

Complexity lower bounds of first-order Riemannian optimization — Open problem

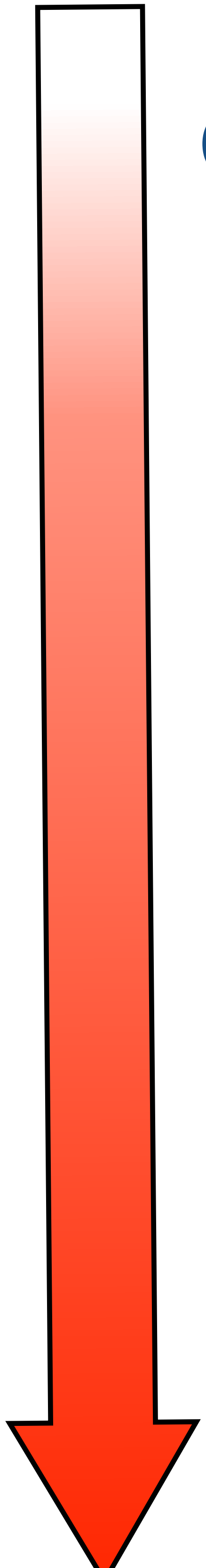
Summary and outlook

Convex

?

Non-convex

Summary and outlook



Convex

?

Non-convex

Beyond Riemannian geometry
Ellipsoid algorithm
Interior point methods
Gradient, stochastic gradient
Accelerated gradient

Summary and outlook

Convex

Beyond Riemannian geometry
Ellipsoid algorithm
Interior point methods
Gradient, stochastic gradient
Accelerated gradient

Local ⇒ global

→ Geometry

Obtain analogs of ellipsoid, IPMs
Recover convex results as special cases as curvature goes to zero
Develop Accelerated gradient
Geometry + neural networks

?

Non-convex

Summary and outlook

Convex

Beyond Riemannian geometry
Ellipsoid algorithm
Interior point methods
Gradient, stochastic gradient
Accelerated gradient

Local ⇒ global

Geometry

Obtain analogs of ellipsoid, IPMs
Recover convex results as special cases as curvature goes to zero
Develop Accelerated gradient
Geometry + neural networks

?

Non-convex

Faster algorithms for large-scale
Interplay of optimization with generalization
When are local optima sufficient?
New models for ML problems
Automating optimization (“ML for OPT”)
Understand SGD!

Thanks!