Optimization for Machine Learning

Lecture 18: Geometric Optimization — II

6.881: MIT

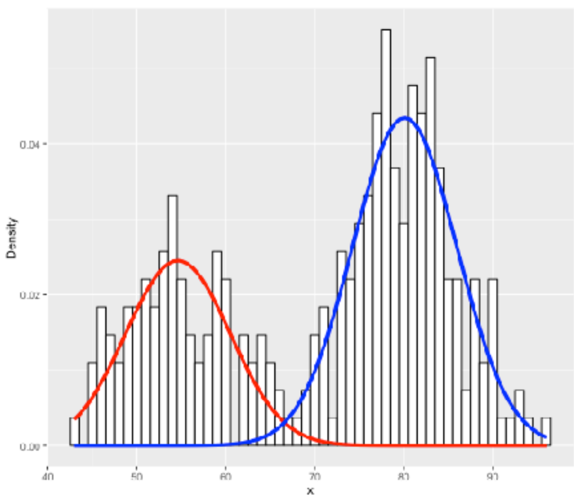
Suvrit Sra
Massachusetts Institute of Technology

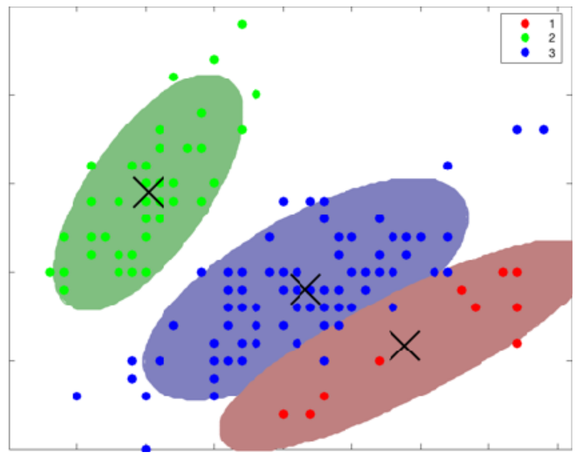
April 29, 2021

Non-convex example

(not g-convex either)

Gaussian mixture models

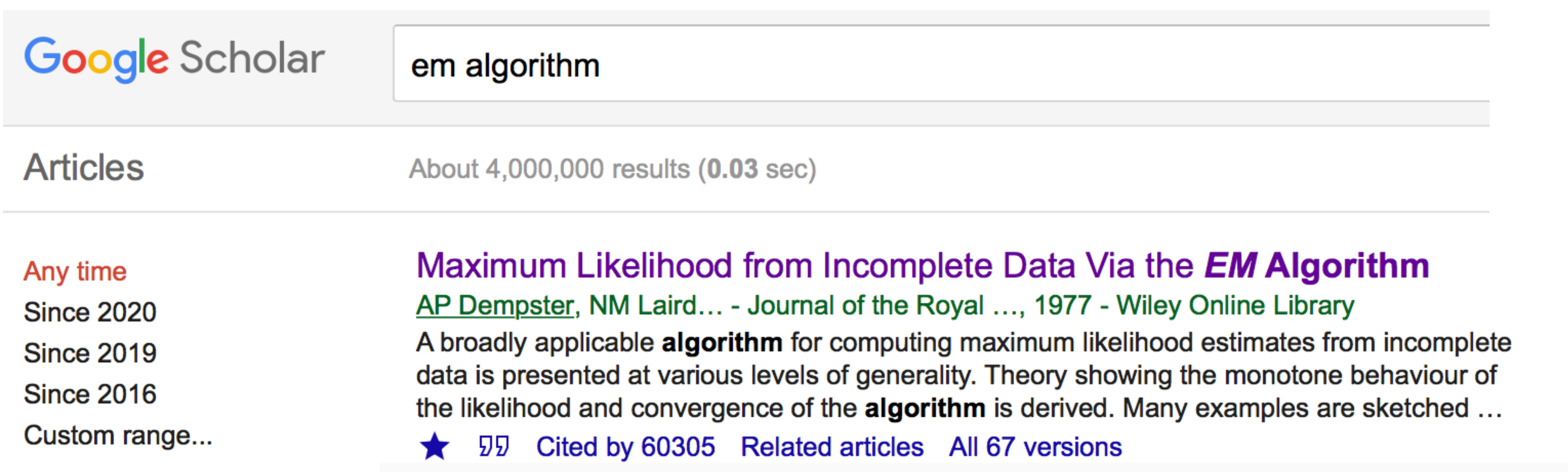




$$p(x) = \sum_k \pi_k \text{Gaussian}(x; \mu_k, \Sigma_k)$$

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

Expectation maximization (EM): the default choice



EM algorithm

Assume $p(x) = \sum_{j=1}^{K} \pi_j p(x; \theta_j)$ is mixture density.

$$\ell(\mathcal{X}; \Theta) := \sum_{i=1}^{n} \ln\Big(\sum_{j=1}^{K} \pi_j p(x_i; \theta_j)\Big).$$

Use convexity of $-\log t$ to compute lower-bound

$$\ell(\mathcal{X}; \Theta) \geq \sum_{ij} \beta_{ij} \ln\left(\pi_j p(x_i; \theta_j)/\beta_{ij}\right).$$

Lecture 13

E-Step:

$$\beta_{ik} = \frac{\pi_k \mathcal{N}(x_i | \Sigma_k)}{\sum_j \pi_j \mathcal{N}(x_i | \Sigma_j)}$$

(generic step, nothing special about Gaussians used here)

EM algorithm

Assume $p(x) = \sum_{j=1}^{K} \pi_j p(x;\theta_j)$ is mixture density.

$$\ell(\mathcal{X};\Theta) := \sum_{i=1}^{n} \ln\Big(\sum_{j=1}^{K} \pi_j p(x_i;\theta_j)\Big).$$

Use convexity of $-\log t$ to compute lower-bound

Lecture 13

$$\ell(\mathcal{X};\Theta) \geq \sum_{ij} \beta_{ij} \ln\left(\pi_j p(x_i;\theta_j)/\beta_{ij}\right).$$

M-step:

$$\max_{\Sigma_1,\ldots,\Sigma_K} \sum_{ij} \beta_{ij} \log\Big(\pi_j \mathcal{N}(x_i|\Sigma_j)/\beta_{ij}\Big)$$

Breaks up into K “weighted” concave MLE problems that admit a closed-form solution, making EM for Gaussians attractive.

$$\Sigma_k = \frac{1}{\sum_i \beta_{ik}} \sum_i \beta_{ik} x_i x_i^T$$

PSD by construction

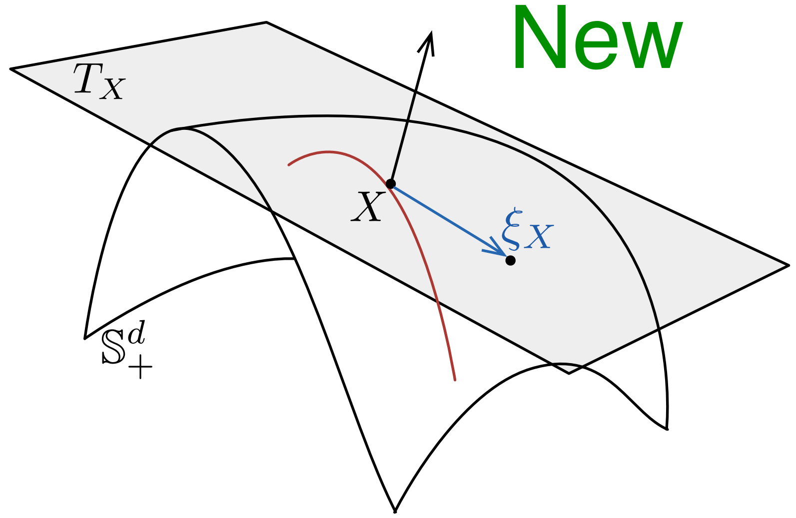
Optimizing GMM log-likelihood

- **Nonconvex –** difficult, possibly several local optima
- **Theory -** Recent progress (Moitra, Valiant 2010; Daskalakis et al, 2017; more!)
- **In Practice –** EM still default: reasons not just "beliefs"!

Key challenge: How to incorporate the positive definiteness constraint on Σ_k

Riemannian

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

Revisiting 1 component MLE

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
(M-step of EM uses this!)
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	Riemannian 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?

An alternative to EM for Gaussian mixture models: batch and stochastic Riemannian optimization

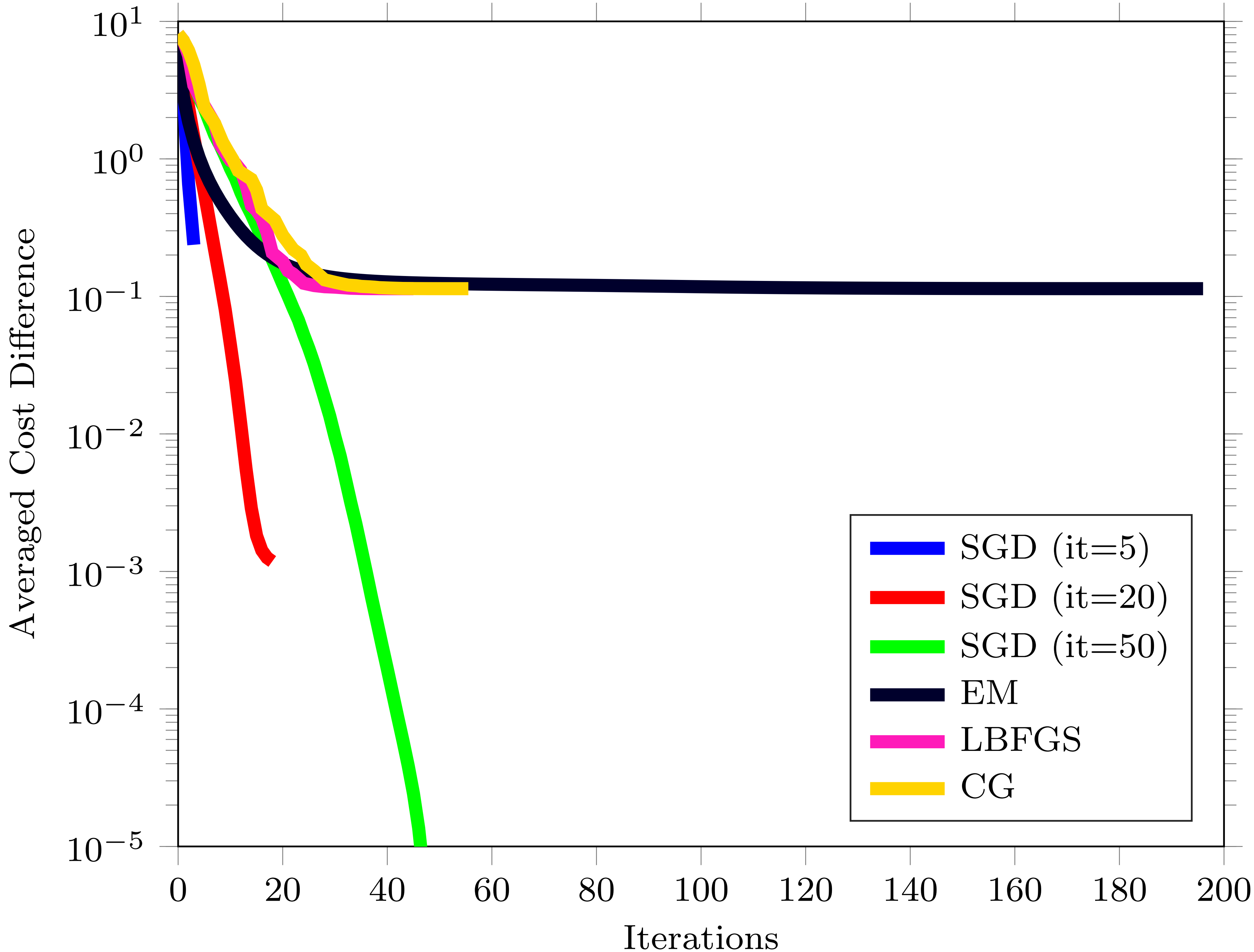
Reshad Hosseini[1,2] · Suvrit Sra[3]

R-LBFGS and Riemannian SGD (without boundedness)

Theorem 4 *Assume a slightly modified version of SGD which output a point x_a by randomly picking one of the iterates, say x_t, with probability $p_t := (2\eta_t - L\eta_t^2)/Z_T$, where $Z_T = \sum_{t=1}^{T}(2\eta_t - L\eta_t^2)$. Furthermore, choose $\eta_t = \min\{L^{-1}, c\sigma^{-1}T^{-1/2}\}$ for a suitable constant c. Then, we obtain the following bound on $\mathbb{E}[\|\nabla f(x_a)\|^2]$, which measures the expected gap to stationarity:*

$$\mathbb{E}[\|\nabla f(x_a)\|^2] \leq \frac{2L\Delta_1}{T} + (c + c^{-1}\Delta_1)\frac{L\sigma}{\sqrt{T}} = \mathcal{O}\left(\frac{1}{T}\right) + \mathcal{O}\left(\frac{1}{\sqrt{T}}\right). \quad (23)$$

Empirical results: Riemannian SGD



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

[Hosseini, Sra, 2017, 2019]

Convergence Theory

G-convex functions: key definitions

$$f(\gamma_{xy}(t)) \equiv f((1-t)x \oplus ty) \leq (1-t)f(x) + tf(y)$$

$$f(x) \geq f(y) + \langle \nabla f(y), \operatorname{Exp}_y^{-1}(x) \rangle_y$$

$$f(x) \geq f(y) + \langle \nabla f(y), \operatorname{Exp}_y^{-1}(x) \rangle_y + \tfrac{\mu}{2} d^2(x,y)$$

Lipschitz continuity

$$|f(x) - f(y)| \leq L_f d(x,y)$$

$$f(x) \leq f(y) + \langle \nabla f(y), \operatorname{Exp}_y^{-1}(x) \rangle_y + \tfrac{L}{2} d^2(x,y)$$

Convergence rate: subgradient method

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

1
$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta g_t - x^*\|^2$$
$$= \|x_t - x^*\|^2 - 2\langle \eta g_t, x_t - x^*\rangle + \eta^2\|g_t\|^2$$
$$\langle -g_t, x^* - x_t\rangle = \frac{1}{2\eta}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] + \frac{\eta}{2}\|g_t\|^2$$

2
$$f(x_t) - f(x^*) \leq \langle -g_t, x^* - x_t\rangle$$ **(convexity)**

3
$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2T\eta}[\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2] + \frac{L_f^2\eta}{2}$$

4
$$\|x_1 - x^*\| \leq D, \eta = D/(L_f\sqrt{T})$$

$$\frac{1}{T}\sum_t f(x_t) - f(x^*) \leq O\left(\frac{1}{\sqrt{T}}\right)$$

Convergence rate: subgradient method

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

1
$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta g_t - x^*\|^2$$
$$= \|x_t - x^*\|^2 - 2\langle \eta g_t, x_t - x^* \rangle + \eta^2 \|g_t\|^2$$
$$\langle -g_t, x^* - x_t \rangle = \frac{1}{2\eta}\left[\|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2\right] + \frac{\eta}{2}\|g_t\|^2$$

2
$$f(x_t) - f(x^*) \leq \langle -g_t, x^* - x_t \rangle$$ **(convexity)**

3
$$\frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \frac{1}{2T\eta}\left[\|x_1 - x^*\|^2 - \|x_{T+1} - x^*\|^2\right] + \frac{L_f^2 \eta}{2}$$

4
$$\|x_1 - x^*\| \leq D, \eta = D/(L_f\sqrt{T})$$

$$\frac{1}{T}\sum_t f(x_t) - f(x^*) \leq O\left(\frac{1}{\sqrt{T}}\right)$$

Convergence rate: Riemannian subgrad

$$x_{t+1} = \mathrm{Exp}_{x_t}(-\eta_t g_t)$$

1

$$d^2(x_{t+1}, x^*)^2 = d^2(\mathrm{Exp}_{x_t}(-\eta g_t), x^*)$$

$$= d^2(x_t, x^*) - ??$$

2

$$f(x_t) - f(x^*) \leq \langle -g_t, \mathrm{Exp}_{x_t}^{-1}(x^*) \rangle$$

(g-convexity)

3

$$\tfrac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*) \leq \tfrac{1}{2T\eta}[d^2(x_1, x^*) - d^2(x_{T+1}, x^*)] + \frac{L_f^2 \zeta \eta}{2}$$

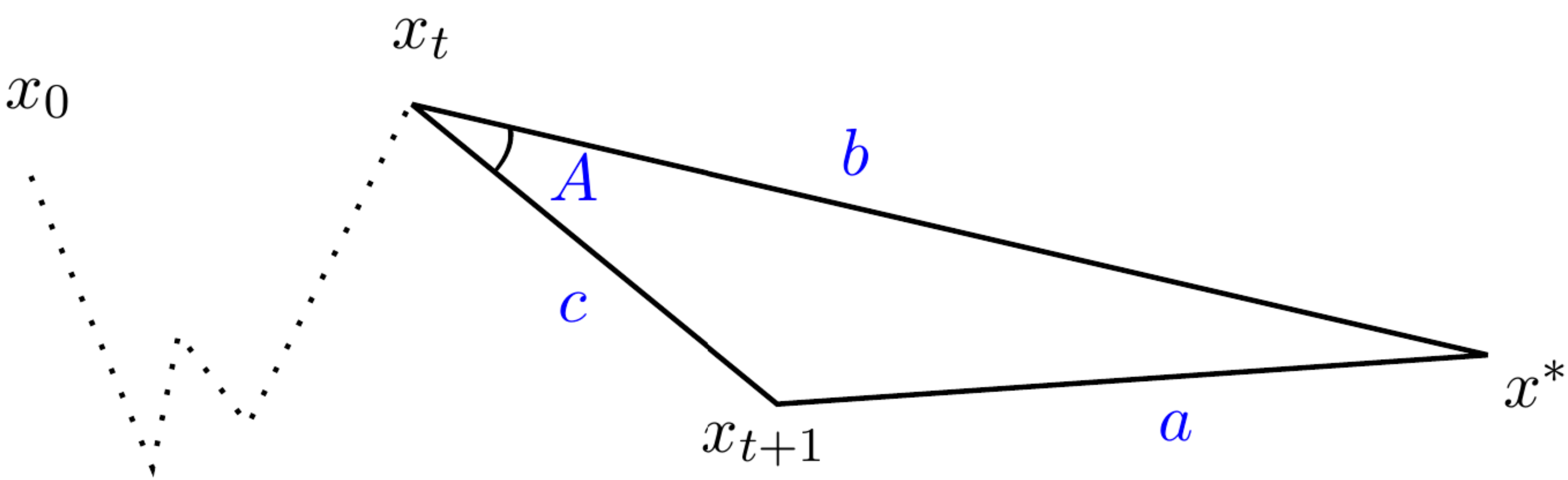
4

$$d(x_1, x^*) \leq D, \eta = D/(L_f\sqrt{\zeta T})$$

$$\tfrac{1}{T}\sum_t f(x_t) - f(x^*) \leq O\left(\sqrt{\tfrac{\zeta}{T}}\right)$$

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

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



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

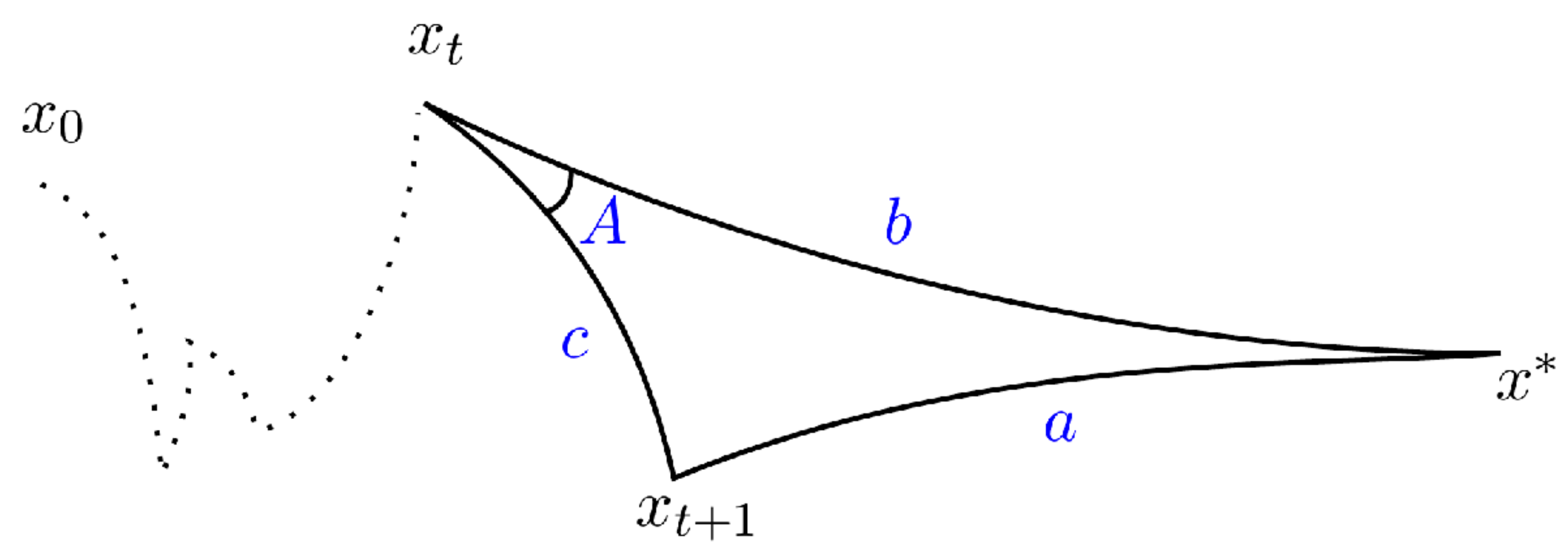
$$\Downarrow$$

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

There's a corresponding **inequality** to bound $d^2(x_{t+1}, x^*)$ on manifolds (and related spaces)

$$x_{t+1} = \mathrm{Exp}_{x_t}(-\eta_t g_t)$$

Based on comparison theorems in Riemannian Geometry



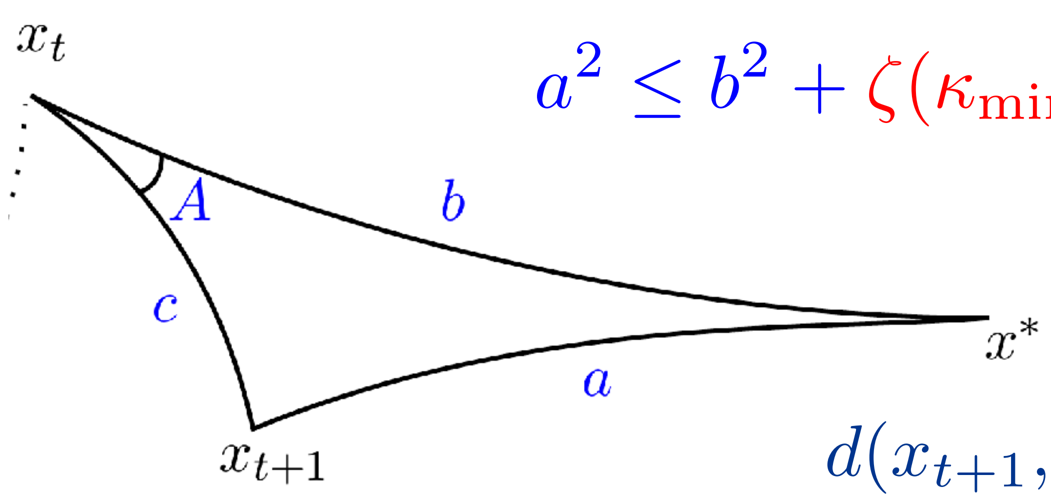
$$a^2 \leq b^2 + \zeta(\kappa_{\min}, b)c^2 - 2bc\cos(A)$$

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

Convergence rate: Riemannian subgrad

$$x_{t+1} = \mathrm{Exp}_{x_t}(-\eta_t g_t) \qquad \textbf{(Riem-subgrad)}$$

1 $$\langle -g_t, \mathrm{Exp}_{x_t}^{-1}(x^*) \le \frac{1}{2\eta}\left[d^2(x_t, x^*) - d^2(x_{t+1}, x^*)\right] + \frac{\zeta(\kappa, d(x_t, x^*))\eta}{2}\|g_s\|^2$$



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

$$d(x_{t+1}, x_t) = \eta\|g_t\|$$

$$d(x_{t+1}, x_t)d(x_t, x^*)\cos(\measuredangle x_{t+1}x_t x^*) = \langle -\eta g_t, \mathrm{Exp}_{x_t}^{-1}(x^*)\rangle$$

4 $$\frac{1}{T}\sum\nolimits_t f(x_t) - f(x^*) \le O\left(\sqrt{\frac{\zeta}{T}}\right)$$

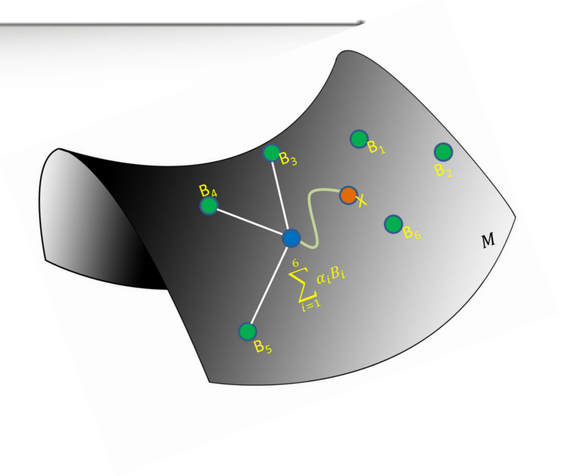
Rates depend on lower bounds on sectional curvature

	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	$O\left(\left(1-\frac{\mu}{L_g}\right)^t\right)$	$O\left(\left(1-\min\left\{\frac{1}{\zeta_{\max}}, \frac{\mu}{L_g}\right\}\right)^t\right)$
Stochastic (sub)gradient	... ...	

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

See paper for other basic 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
- Riemannian SPIDER (optimal rates)

[Zhang, Reddi, Sra, NIPS 2016] *[Zhang, Zhang, Sra, 2018]*

Stochastic Optimization

$$\min_{x \in \mathcal{M}} \quad f(x) = \mathbb{E}[F(x,\xi)]$$

Fast stochastic optimization on Riemannian Manifolds
Hongyi Zhang, Sashank Reddi, Suvrit Sra.
NIPS 2016.

R-SPIDER: A Fast Riemannian Stochastic Optimization Algorithm with Curvature Independent Rate
Jingzhao Zhang, Hongyi Zhang, Suvrit Sra.
arXiv:1811.04194

Optimal rates for g-convex still open

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

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

Open problem

Accelerated gradient

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

From Nesterov's Estimate Sequence to Riemannian Acceleration
Kwangjun Ahn, Suvrit Sra
33rd Annual Conference on Learning Theory (COLT'20)

$$
\begin{aligned}
x_{t+1} &\leftarrow y_t + \alpha_{t+1}(z_t - y_t)\\
y_{t+1} &\leftarrow x_{t+1} - \gamma_{t+1}\nabla f(x_{t+1})\\
z_{t+1} &\leftarrow x_{t+1} + \beta_{t+1}(z_t - x_{t+1}) - \eta_{t+1}\nabla f(x_{t+1})
\end{aligned}
$$

Nesterov's AGM

Riemannnian AGM

$$
\begin{aligned}
x_{t+1} &\leftarrow \mathrm{Exp}_{y_t}\left(\alpha_{t+1}\mathrm{Exp}_{y_t}^{-1}(z_t)\right)\\
y_{t+1} &\leftarrow \mathrm{Exp}_{x_{t+1}}\left(-\gamma_{t+1}\nabla f(x_{t+1})\right)\\
z_{t+1} &\leftarrow \mathrm{Exp}_{x_{t+1}}\left(\beta_{t+1}\mathrm{Exp}_{x_{t+1}}^{-1}(z_t) - \eta_{t+1}\nabla f(x_{t+1})\right)
\end{aligned}
$$

Accelerated gradient

From Nesterov's Estimate Sequence to Riemannian Acceleration
Kwangjun Ahn, Suvrit Sra
33rd Annual Conference on Learning Theory (COLT'20)

Theorem 1.1 (Informal) *Let f be L-smooth and μ-strongly convex in a geodesic sense. Then, there exists a* computationally tractable *optimization algorithm satisfying*

$$f(x_t) - f(x_*) = O\left((1-\xi_1)(1-\xi_2)\cdots(1-\xi_t)\right),$$

*where $\{\xi_t\}$ satisfies (i) $\{\xi_t\}_{t\geq 1} > {}^{\mu}/_{L}$ (**strictly faster** than gradient descent); and (ii) $\exists\lambda \in (0,1)$ such that $\forall\epsilon > 0$, $|\xi_t - \sqrt{{}^{\mu}/_{L}}| \leq \epsilon$, for $t \geq \Omega\left(\frac{\log(1/\epsilon)}{\log(1/\lambda)}\right)$ (eventually achieves **full acceleration**).*

Challenge: deciding what ξ_t should be, remaining implementable

Riemannian Frank-Wolfe

Riemannian Frank-Wolfe and Stochastic Frank-Wolfe Methods
Melanie Weber, Suvrit Sra
arXiv:1910.04194, arXiv:1710.10770

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

$$\text{s.t.} \quad x \in \mathcal{X}$$

Projection-free methods for constrained optimization

(involves non-convex subproblems though)

Some other works

Proximal point method for vector optimization on Hadamard manifolds
Glaydston de C.Bento, Orizon P.Ferreira, Yuri R.L.Pereira

What do `convexities' imply on Hadamard manifolds?
Alexandru Kristály, Chong Li, Genaro Lopez, Adriana Nicolae

Convex Analysis and Optimization in Hadamard Spaces
Miroslav Bacak, 2014 de Gruyter Publishers

Global rates of convergence for nonconvex optimization on manifolds
Nicolas Boumal, P-A Absil, Coralia Cartis

Averaging Stochastic Gradient Descent on Riemannian Manifolds
Nilesh Tripuraneni, Nicolas Flammarion, Francis Bach, Michael I. Jordan

Optimization Techniques on Riemannian Manifolds
Steven Thomas Smith

…and many others