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
- **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)
Example: Riemannian optimization

- Orthogonality constraint
- Fixed-rank constraint
- Positive semi-definite constraint

- Stiefel manifold
- Grassmann manifold
- PSD manifold

Vector space optimization ↔ Riemannian optimization

[Udriste, 1994; Absil et al., 2009]
Function classes of interest

- Convex
- Lipschitz
- Smooth
- Strongly convex
Function classes of interest

- Convex
- Lipschitz
- Smooth
- Strongly convex

Geodesically

$T_x M$

$y(t)$
What is geodesic convexity?

Convexity

Geodesic convexity

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

on riemannian manifold

\[ f(y) \geq f(x) + \langle g_x, \operatorname{Exp}_x^{-1}(y) \rangle_x \]

**Metric spaces & curvature:** [Menger; Alexandrov; Busemann; Bridson, Häflinger; Gromov; Perelman]
Positive definite matrix manifold

Geodesic

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

Examples

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

Verify

\[ f(X \#_t Y) \leq (1 - t)f(X) + tf(Y) \]
Recognizing, constructing, and optimizing g-convex functions

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

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

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

(jointly g-convex)

Many more theorems and corollaries

One-D version known as: Geometric Programming

www.stanford.edu/~boyd/papers/gp_tutorial.html

Examples

\[ X \succeq 0 \]
Matrix square root

Broadly applicable

Key to ‘expm’, ‘logm’
Matrix square root

Nonconvex optimization through the Euclidean lens

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

**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
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

\[
\min_{X \succ 0} \quad \delta^2_S(X, A) + \delta^2_S(X, I)
\]

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

\[
\delta^2_S(X, Y) := \frac{1}{2} \log \det \left( \frac{X + Y}{2} \right) - \frac{1}{2} \log \det(XY)
\]
Matrix square root

50 × 50 matrix $I + \beta UU^T$

$\kappa \approx 64$
Brascamp-Lieb Constant

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

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

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

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

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

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

**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

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

Goal

given pairwise constraints learn Mahalanobis distance

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

Positive definite matrix \( A \)
Metric learning methods

**MMC**
[Xing, Jordan, Russell, Ng 2002]

**LMNN**
[Weinberger, Saul 2005]

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

**tons of other methods!**
A simple new way for metric learning

Euclidean idea

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

New idea

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

Equivalently solve

\[
\min_{A > 0} h(A) := \text{tr}(AS) + \text{tr}(A^{-1}D)
\]

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

\[
D := \sum_{(x_i, x_j) \in D} (x_i - x_j)(x_i - x_j)^T
\]
A simple new way for metric learning

Closed form solution!

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

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

More generally

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

\[ S^{-1} \#_t D \]
## Experiments

<table>
<thead>
<tr>
<th>Data Set</th>
<th>GMML</th>
<th>LMNN</th>
<th>ITML</th>
<th>FlatGeo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Segment</td>
<td>0.0054</td>
<td>77.595</td>
<td>0.511</td>
<td>63.074</td>
</tr>
<tr>
<td>Letters</td>
<td>0.0137</td>
<td>401.90</td>
<td>7.053</td>
<td>13543</td>
</tr>
<tr>
<td>USPS</td>
<td>0.1166</td>
<td>811.2</td>
<td>16.393</td>
<td>17424</td>
</tr>
<tr>
<td>Isolet</td>
<td>1.4021</td>
<td>3331.9</td>
<td>1667.5</td>
<td>24855</td>
</tr>
<tr>
<td>MNIST</td>
<td>1.6795</td>
<td>1396.4</td>
<td>1739.4</td>
<td>26640</td>
</tr>
</tbody>
</table>

[Habibzadeh, Hosseini, Sra ICML 2016]
Gaussian mixture models

\[ p_{\text{mix}}(x) := \sum_{k=1}^{K} \pi_k p_N(x; \Sigma_k, \mu_k) \]

\[ \max \prod_i p_{\text{mix}}(x_i) \]

Expectation maximization (EM): default choice

\[ p_N(x; \Sigma, \mu) \propto \frac{1}{\sqrt{\det(\Sigma)}} \exp \left( -\frac{1}{2} (x - \mu)^T \Sigma^{-1} (x - \mu) \right) \]

[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 $\Sigma_k$

Geometric opt  
Unconstrained, Cholesky

New  
Folklore

\[ LL^T \]
## Failure of geometric optimization

<table>
<thead>
<tr>
<th>$K$</th>
<th>EM</th>
<th>Riem-CG</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17s / 29.28</td>
<td>947s / 29.28</td>
</tr>
<tr>
<td>5</td>
<td>202s / 32.07</td>
<td>5262s / 32.07</td>
</tr>
<tr>
<td>10</td>
<td>2159s / 33.05</td>
<td>17712s / 33.03</td>
</tr>
</tbody>
</table>

$manopt.org$

Riemannian opt. toolbox

d=35 images dataset
## Failure of “obvious” $LL^T$

<table>
<thead>
<tr>
<th>sep.</th>
<th>EM</th>
<th>CG-LL$^T$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2</td>
<td>52s // 12.7</td>
<td>614s // 12.7</td>
</tr>
<tr>
<td>1</td>
<td>160s // 13.4</td>
<td>435s // 13.5</td>
</tr>
<tr>
<td>5</td>
<td>72s // 12.8</td>
<td>426s // 12.8</td>
</tr>
</tbody>
</table>

\[
\| \mu_i - \mu_j \| \geq \text{sep} \max_{i,j} \{ \text{tr} \Sigma_i, \text{tr} \Sigma_j \} \quad \text{d=20 simulation}
\]
What's wrong?

log-likelihood for one component

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

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

Euclidean convex problem

**Not** geodesically convex

Mismatched geometry?
Reformulate as $g$-convex

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

$$\max_{S > 0} \hat{\mathcal{L}}(S) := \sum_{i=1}^{n} \log q_N(y_i; S),$$

**Thm.** The modified log-likelihood is $g$-convex. Local max of modified mixture LL is local max of original.
### Success of geometric optimization

<table>
<thead>
<tr>
<th>K</th>
<th>EM</th>
<th>Riem-CG</th>
<th>L-RBFGS</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>17s // 29.28</td>
<td>18s // 29.28</td>
<td>14s // 29.28</td>
</tr>
<tr>
<td>5</td>
<td>202s // 32.07</td>
<td>140s // 32.07</td>
<td>117s // 32.07</td>
</tr>
<tr>
<td>10</td>
<td>2159s // 33.05</td>
<td>1048s // 33.06</td>
<td>658s // 33.06</td>
</tr>
</tbody>
</table>

---

**Riem-CG (manopt) savings:**

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

<table>
<thead>
<tr>
<th>d=35 images dataset</th>
</tr>
</thead>
</table>

[github.com/utvisionlab/mixest](https://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

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

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

oracle access to exact or stochastic (sub)gradients

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

analog to: \( x \leftarrow x - \eta \nabla f(x) \)
In particular, we study the **global complexity of first-order g-convex optimization**

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

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

Grönwall's inequality

Toponogov's theorem

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

\[
\zeta(\kappa_{\text{min}}, b) \triangleq \frac{\sqrt{|\kappa_{\text{min}}|b}}{\tanh(\sqrt{|\kappa_{\text{min}}|b})}
\]
Convergence rates depend on lower bounds on the sectional curvature

(Sub)gradient

<table>
<thead>
<tr>
<th>Convex</th>
<th>g-convex</th>
</tr>
</thead>
<tbody>
<tr>
<td>(O(\sqrt{\frac{1}{t}}))</td>
<td>(O(\sqrt{\frac{\zeta_{\text{max}}}{t}}))</td>
</tr>
<tr>
<td>(O(\frac{1}{t}))</td>
<td>(O(\frac{\zeta_{\text{max}}}{t}))</td>
</tr>
<tr>
<td>(O\left((1 - \frac{\mu}{L_g})^t\right))</td>
<td>(O\left((1 - \min\left{\frac{1}{\zeta_{\text{max}}}, \frac{\mu}{L_g}\right})^t\right))</td>
</tr>
</tbody>
</table>

Stochastic (sub)gradient

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

See paper for other interesting results [Zhang, Sra, COLT 2016]
G-nonconvex optimization

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

- \(\mathcal{M}\) 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]