DS-607 - Optimization for Data Science

- focus on optimization algorithms over non-Euclidean geometry
- first half will be covering optimization over smooth manifolds
- At some point in second half, we will start computational optimal transport.
- Aim: Understand the concepts, read papers, implement algorithms, propose novel solutions for various applications

- Pre-requisites include some understanding of. linear algebra, optimization, probability & statistics and ML.

- References (provided in course logistics page)

- Attendance is compulsory.

- Grading: some combination of assignments, quizes, midsem, endsem, course project.

- Scribing?

Consider the following problem:

$$\min_{x \in S} \quad f(x) \qquad\qquad f: S \to \mathbb{R}$$

What informations will influence our algorithms

Say $S = \mathbb{R}^n$, i.e., a linear space.

$$\min_{x \in \mathbb{R}^n} f(x)$$

$\nabla f(x)$ $\quad \langle \cdot, \cdot \rangle$

$\frac{1}{2} \nabla f(x)$ $\quad 2\langle \cdot, \cdot \rangle$

Say $f(\cdot)$ is differentiable and we apply the standard gradient descent

What implicit choice are we making?

The choice of inner product $\langle u, v \rangle = 2u^T v$

We choose to turn $\mathbb{R}^n$ into a Euclidean space endowed with the standard inner product.

Say $S = \mathcal{M}$ is a smooth manifold.

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

We choose to turn $\mathcal{M}$ into a Riemannian manifold.

If $f(\cdot)$ is smooth, we have a Riemannian gradient and can use it to develop Riemannian gradient descent algorithm.

Example of Optimization problem on manifolds.

eg.) $A \in \mathbb{R}^{n \times n}$, $A = A^T$. Obtain largest eigenvalue of A?

If $A = A^T$, $A = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \in \mathbb{R}$

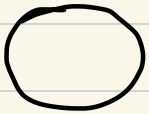
$v_1, \cdots v_n$: orthogonal

$\|v_i\| = 1$, $v_i^T v_j = 0$

$f(x) = x^T A x$

$\max_{x \in \mathbb{R}^n \text{ s.t. } \|x\|=1} x^T A x = \lambda_1$

$S^{n-1} = \{x \in \mathbb{R}^n \mid x^T x = 1\}$ is a manifold.



Eg: $A \in \mathbb{R}^{m \times n}$. Obtain largest singular value of A

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^T \quad \text{(SVD)} \qquad \sigma_1 \geq \cdots \geq \sigma_r$$

$u_1, \ldots u_r$ are orthonormal in $\mathbb{R}^m$

$v_1, \ldots v_r$ are " " $\mathbb{R}^n$

$$f(x, y) = x^T A y$$

$$\max_{\substack{x \in \mathbb{R}^m, y \in \mathbb{R}^n \\ \|x\|=1, \|y\|=1}} x^T A y = \sigma_1$$

$S^{n-1} \times S^{m-1}$ is manifold

eg: $A = A^T$, $A = A^T$, want top-k eigenspace of A.

Consider the problem

$f(X) = \mathrm{Tr}(X^T A X)$ where $X \in St(n,k) = \{X \in \mathbb{R}^{n \times k} \mid X^T X = I_k\}$

Stiefel manifold.

$$\max_{\substack{X \in \mathbb{R}^{n \times k} \\ X^T X = I_k}} f(X) = \lambda_1 + \lambda_2 + \cdots + \lambda_k.$$

Let $Q \in O(k) = \{ Q \in \mathbb{R}^{k \times k} \mid Q^T Q = Q Q^T = I_k \}$

Then we note that

$$f(X) = \mathrm{Tr}(X^T A X) \qquad X \in St(n,k)$$

$$= \mathrm{Tr}((XQ)^T A (XQ)) \,, \quad X \in St(n,k),\ Q \in O(k)$$

$$= \mathrm{Tr}(Q^T X^T A X Q)$$

$$= \mathrm{Tr}(X^T A X)$$

$O(k)$ is a manifold.

Define: $X_1^T X_1 = I_k$, $X_2^T X_2 = I_k$

for $X_1, X_2 \in St(n,k)$, $X_1 \sim X_2 \iff X_1 = X_2 Q$ for some $Q \in O(k)$

What is common between X_1 and X_2 s.t. $X_1 \sim X_2$

other than $X_1 \in St(n,k)$ and $X_2 \in St(n,k)$?

$span(X_1) = span(X_2)$

Let the equivalence class corresponding to $X \in St(n,k)$ be denoted by $[X]$

Then, $[X] \in St(n,k)/\sim$ (Stiefel manifold modulo equivalence relation)

Visualize $St(n,k)$ as all possible ways of choosing a co-ordinate system for a k-dimensional subspace in $\mathbb{R}^n$

Then $St(n,k)/\mathcal{O}$ is just the set of all the planes ignoring the co-ordinate system

$St(n,k)/\mathcal{O}$ is called the quotient space and is a manifold.

Applications

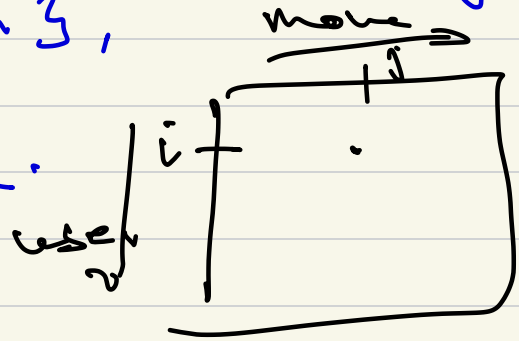
1. Low-rank matrix learning:

The set $\mathbb{R}_r^{m \times n} = \{X \in \mathbb{R}^{m \times n} : \text{rank}(X) = r\}$ is a manifold.

- Netflix Challenge: Given 'm' users and 'n' movies,

X_{ij} representing the score (1-5) that a user i gives to movie j, and $\Omega = \{(i,j) \mid \text{score given by user } i \text{ to movie } j \text{ is known}\}$,

Complete. X given X_Ω.

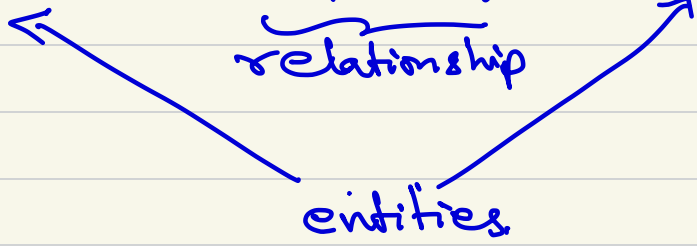


- Image or video completion - few pixels or frames are corrupted.

- Link prediction in knowledge graphs / social networks

Discover new facts of form subject-predicate-object.
given a few such facts

eg. { Mumbai, capital-of, Maharashtra }



These are modeled using low-rank tensors

2. Symmetric Positive definite matrices:

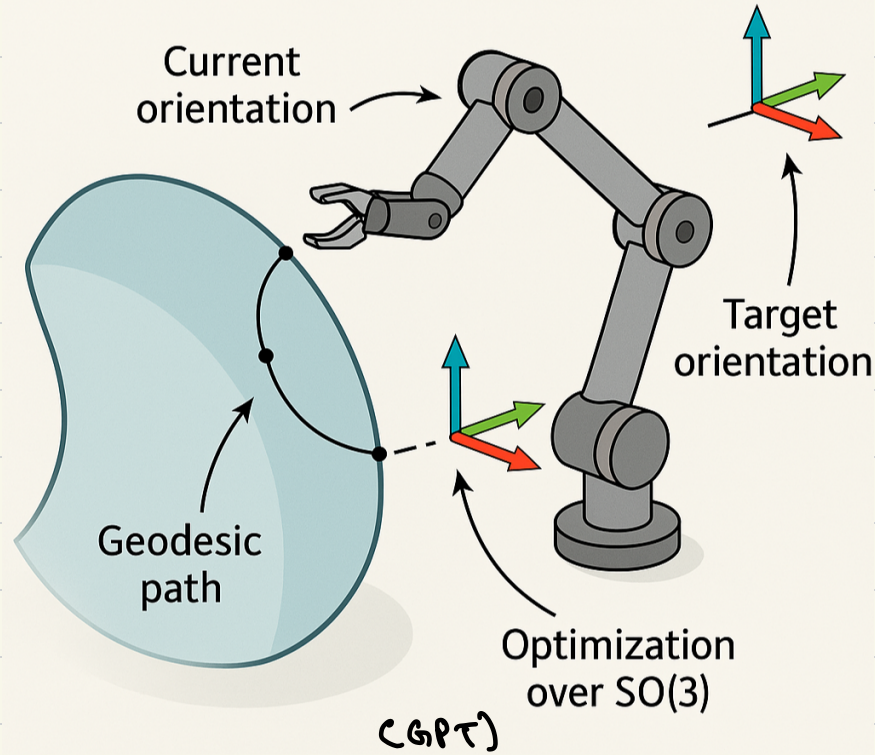
$$\{ S \in \mathbb{R}^{d \times d} \mid S \succ 0 \}$$

Used for modeling covariance matrices

- Modeling mixture of Gaussians
- Classifying radar data, EEG recordings, or human action (using, e.g., SPDNet)

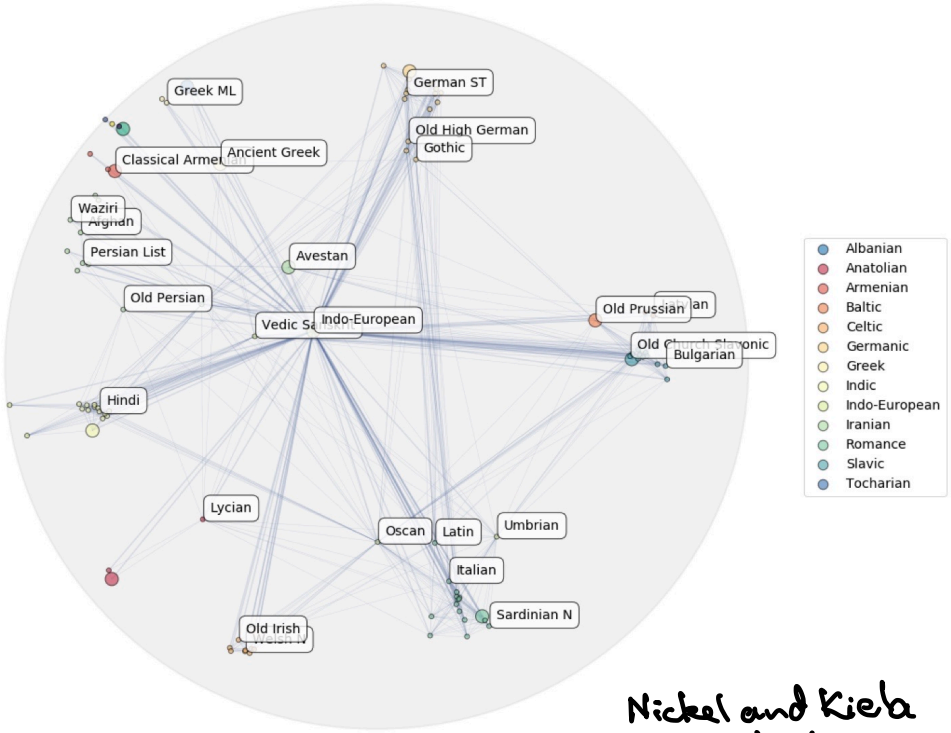
- Raw motion data (from video or sensors)
- features include joint positions, velocities over time
- Covariance matrix X is computed.
- Classify $\phi(X)$ ($\succ 0$) obtained via SPD-Net

(3) Orientation learning via SO(3)
(set of rotation matrices)



(GPT)

(4) Hyperbolic spaces - representing hierarchies



Nickel and Kiela et al 2018

Figure 3: Embedding of the IELex lexical cognate data.

Hide & Seek: Transformer Symmetries Obscure Sharpness & Riemannian Geometry Finds It

Marvin F. da Silva [1 2] Felix Dangel [2] Sageev Oore [1 2]

Efficient Optimization with Orthogonality Constraint: a Randomized Riemannian Submanifold Method

Andi Han [1 2] Pierre-Louis Poirion [1] Akiko Takeda [1 3]

Score-based Pullback Riemannian Geometry: Extracting the Data Manifold Geometry using Anisotropic Flows

Willem Diepeveen [*1] Georgios Batzolis [*2] Zakhar Shumaylov [3] Carola-Bibiane Schönlieb [3]

Fast, Accurate Manifold Denoising by Tunneling Riemannian Optimization

Shiyu Wang [1 2] Mariam Avagyan [1 2] Yihan Shen [3 2] Arnaud Lamy [1 2] Tingran Wang [1 2] Szabolcs Márka [4 2] Zsuzsa Márka [5 2] John Wright [1 6 2]

Preconditioned Riemannian Gradient Descent Algorithm for Low-Multilinear-Rank Tensor Completion

Yuanwei Zhang [1] Fengmiao Bian [2] Xiaoqun Zhang [1 3] Jian-Feng Cai [2]

Some recent works in ICML'25 using Riemanian optimization

Lets start with unconstrained convex minimization problems

$$\min_{x \in \mathbb{R}^n} f(x)$$

Assume $f \in C^0(\mathbb{R}^n)$

first order necessary conditions of optimality?

Now, we consider convex minimization problem over convex sets constrained.

$$\min_{x \in \Omega} f(x)$$

$\Omega \subset \mathbb{R}^d$ non-empty, closed, convex

first order necessary conditions for optimality?

Assume x^* is a local soln.

Then

$$\langle \nabla f(x^*), x - x^* \rangle \geq 0$$

$\forall\ x \in \Omega$.

Equivalent way:

$$\langle \nabla f(x^*), p \rangle \geq 0$$

$\forall\ p \in T_\Omega(x^*)$, where

$T_\Omega(x) = \text{closure}\{\lambda(y - x) \mid y \in \Omega, \lambda \geq 0\}$

Let us have a more concrete problem:

$$\min_{x \in \mathbb{R}^d} f(x) \quad \text{s.t} \quad h(x) = 0$$

(eg $x_1 + x_2$) (eg. $x_1^2 + x_2^2 - 1 = 0$)

K.K.T conditions of optimality (necessary)
(for x^* to be a local optimizer)

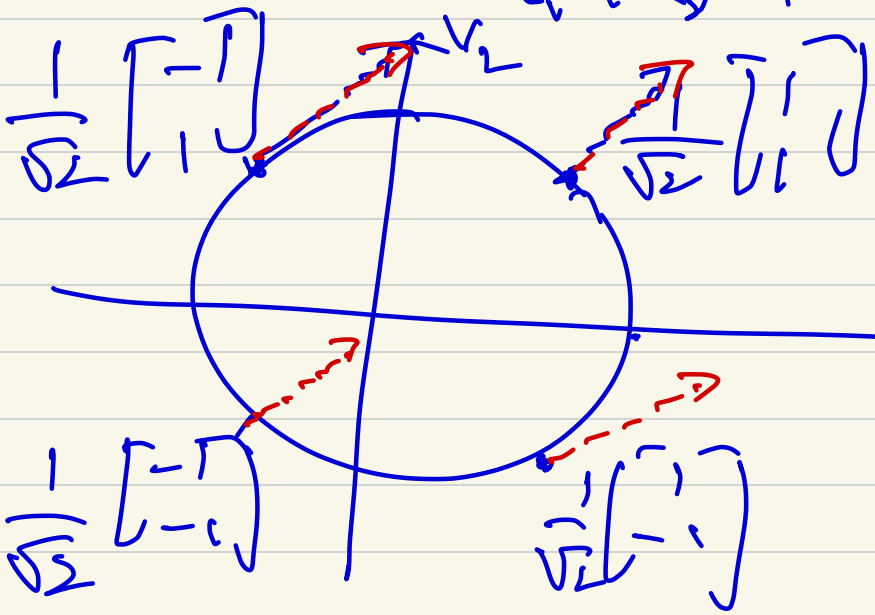
(1) $\nabla f(x) - \lambda \nabla h(x) = 0$

(2) $h(x) = 0$

for some $(x, \lambda) = (x^*, \lambda^*)$

$f(x) = x_1 + x_2$

$h(x) = x_1^2 + x_2^2 - 1$



$\nabla f(x) = \begin{bmatrix}1\\1\end{bmatrix}$

$\nabla h(x) = \begin{bmatrix}2x_1\\2x_2\end{bmatrix}$

$\begin{bmatrix}1\\1\end{bmatrix} - \lambda 2\begin{bmatrix}x_1\\x_2\end{bmatrix} = 0$

$\begin{bmatrix}x_1\\x_2\end{bmatrix} = \frac{1}{2\lambda}\begin{bmatrix}1\\1\end{bmatrix}$

first order analysis

- $h(x) = 0$

$$h(x+s) \simeq h(x) + \nabla h(x)^T s = 0$$

$$\nabla h(x)^T s = 0$$

$$f(x+s) \simeq f(x) + \nabla f(x)^T s$$

If $f(x)$ is locally optimal

$$\Rightarrow \quad \nabla f(x)^T s \geq 0$$

$$\nabla h(x)^T s = 0$$

$$\nabla f(x)^T s < 0$$