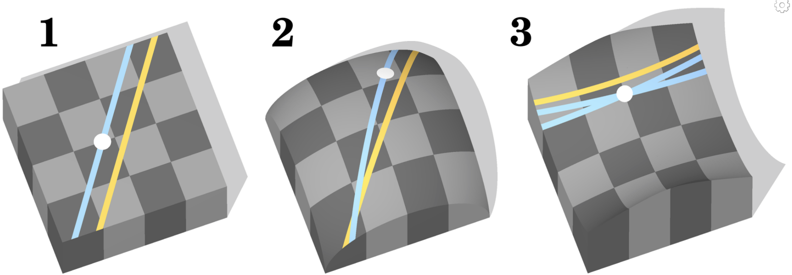
Gradient Descent Algorithms in Hyperbolic Space

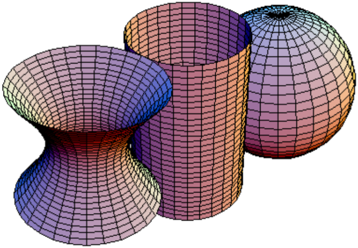
Michael Skinner and Siddartha Devic

Optimization in ML (CS 6301.012), UT Dallas

Taxonomy of Geometries



(1) Euclidean, (2) Elliptical, (3) Hyperbolic



Related Work

- Continuous analog of trees, used in representing WordNet hierarchy space [Nickel and Kiela, 2017].
- Most use alternate representations of hyperbolic space, but [Wilson and Leimeister, 2018] argue that we can perform GD directly in hyperbolic space.
- From the mathematics side, [Bonnabel, 2013] show that SGD generalizes to arbitrary Riemannian manifolds (including $\mathbb{H}^n$).
- Accelerated Riemannian GD [Zhang and Sra, 2018] (COLT '18).
  - Generalize AGD to Riemannian manifolds with convergence bounds (we found out about this paper yesterday).

Methods

- We extend the results in [Wilson and Leimeister, 2018]:
  - Replicate experiments for vanilla GD from [Wilson and Leimeister, 2018]
  - Implement accelerated GD and Barzelia-Borwein for the barycentre problem
  - Implement Armijo backtracking search for selecting learning rate
- Utilize barycentre problem implementation for hyperbolic k-means clustering.
- No euclidean analog to optimization procedure. I.E. we cannot solve the problem in euclidean space and somehow project back.

Background

- [Wilson and Leimeister, 2018] give derivations for GD within $\mathbb{H}^n$ directly.

1. $\Theta \in \mathbb{H}^n$ = current value of the centroid
2. Gradient in the $(n+1)$-dimensional ambient space with respect to one of the arguments of the function measuring the distance between two points has the form:

$$\nabla_u^{\mathbb{R}^{n:1}} d_{\mathbb{H}^n}(u, v) = -((\langle u, v\rangle_{n:1}^2 - 1)^{-\frac{1}{2}} \cdot v.$$

3. Note that $\langle \cdot, \cdot \rangle_{n:1}$ is a special bilinear form in the ambient space defined as

$$\langle \mathbf{u}, \mathbf{v}\rangle_{n:1} = u_1 v_1 + \cdots + u_{n-1} v_{n-1} - u_n v_n \text{ for } \mathbf{u}, \mathbf{v} \in \mathbb{R}^n.$$

Background cont.

4 This gradient is then projected into the tangent space by the following expression:

$$\nabla_{\Theta}^{\mathbb{H}^n} E = \nabla_{\Theta}^{\mathbb{R}^{n:1}} E + \left\langle \Theta, \nabla_{\Theta}^{\mathbb{R}^{n:1}} E \right\rangle_{n:1} \cdot \Theta.$$

5 Finally, the parameter update equation is:

$$\Theta^{new} = \mathrm{Exp}_{\Theta}(-\alpha \cdot \nabla_{\Theta}^{\mathbb{H}^n} E).$$

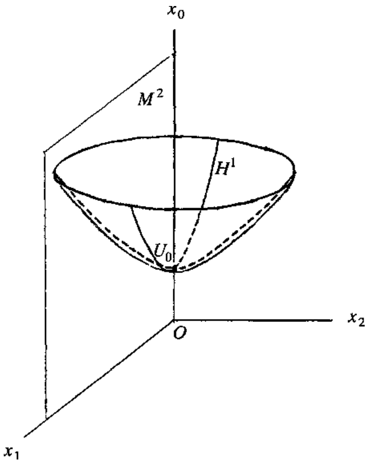
Where Exp_p, the exponential map from the tangent space back to the hyperbolic manifold, is:

$$\mathrm{Exp}_p(v) = \cosh(||v||)p + \sinh(||v||)\frac{v}{||v||}.$$

Example

As an example, consider the hyperboloid $\mathbb{H}^2$ as follows.

$$\mathbb{H}^2 = \{\mathbf{x} \in \mathbb{R}^3 | x_1^2 + x_2^2 - x_3^2 = -1, x_3 > 0\}$$



Results: Vanilla GD

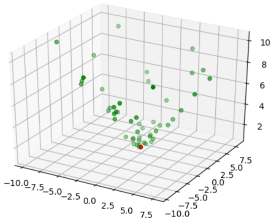


Figure 1: Sampled points on $\mathbb{H}^2$ sitting in $\mathbb{R}^3$ with Karcher mean (red). Karcher mean of $\mathbb{H}^2$ is not the same as centroid in $\mathbb{R}^3$.

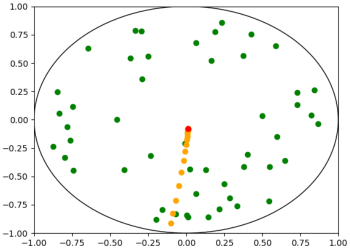


Figure 2: Same sampled points shown in $\mathbb{R}^2$ using Poincaré projection. Path to Karcher mean during GD marked in orange.

Results: Vanilla GD

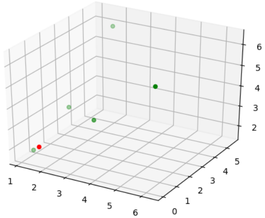


Figure 3: Points in first quadrant of the hyperboloid $\mathbb{H}^2$ depicted with their Karcher mean.

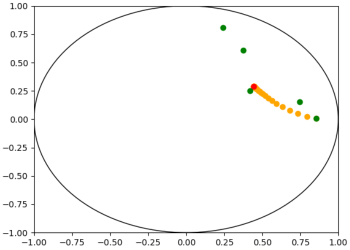


Figure 4: Poincaré projection of points in first quadrant (left). Initializing GD at a point in the point set shows us that the shortest line between two points lies on a geodesic connecting the points.

Results: Vanilla GD

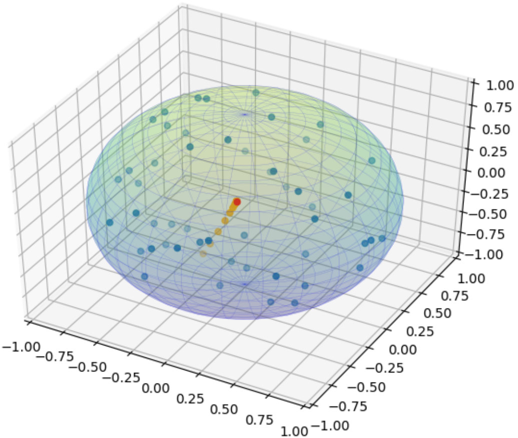


Figure 5: The Poincaré ball projection of $\mathbb{H}^3$ showing convergence of vanilla gradient descent.

Results: Accelerated GD

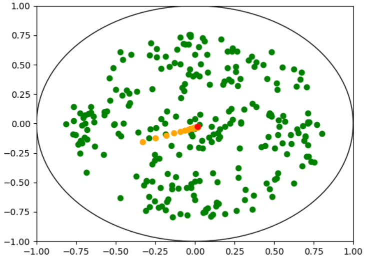


Figure 6: Vanilla gradient descent path ($\alpha = 0.1$), 27 iterations.

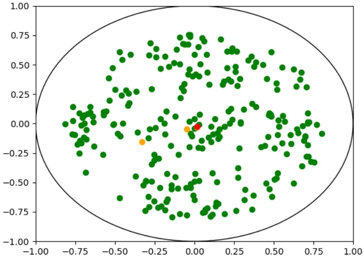
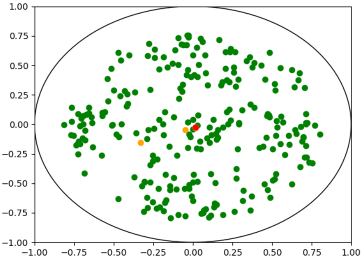
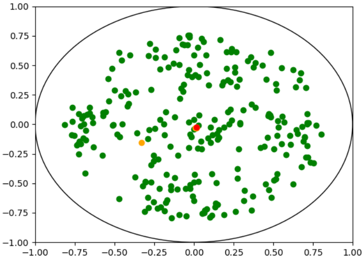


Figure 7: Vanilla GD with Armijo ($\gamma = 1e - 4$), 3 iterations.

Results: Accelerated GD



Accelerated GD with Armijo ($\gamma = 1e - 4$), 6 iterations.



Barzelia-Borwein GD with Armijo ($\gamma = 1e - 4$), 6 iterations.

Results: AGD in Higher Dimensions

- Not straightforward to get working correctly.
- Not clear if convergence guarantees hold, since algorithm is different from the usual GD parameter update step.
- Algorithm in [Zhang and Sra, 2018] may work if we have time to implement.

Clustering

- Barycenter problem is same as centroid update step in k-means
- random, k-means++ init scheme
- Not clear if k-means++ is still $\Theta(\log k)$ competitive.

Clustering

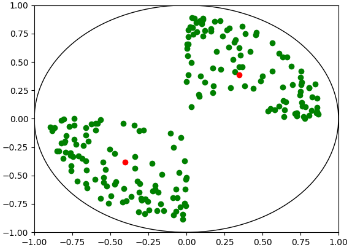


Figure 8: Randomly sampling points in positive and negative octant of $\mathbb{R}^3$ and generating points in $\mathbb{H}^2$ with them.

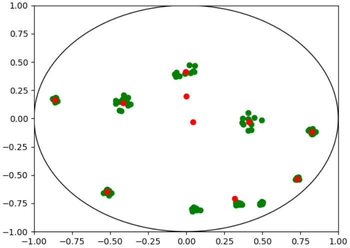


Figure 9: Randomly sample 10 points to serve as centers for hyperbolic balls in $\mathbb{H}^2$, then sample 10 additional points from within each hyperbolic ball (radius $\epsilon = 0.2$).

Clustering Initialization

Dimension	k (Number of clusters)	random init	k++ init
4*5	5	496.23	**494.89**
	10	**403.21**	409.37
	15	379.17	**364.62**
	20	366.44	**337.84**
4*10	5	771.52	**769.43**
	10	**744.82**	748.81
	15	727.83	**705.44**
	20	689.47	**652.09**
4*15	5	951.74	**946.49**
	10	929.26	**920.48**
	15	884.06	**853.19**
	20	842.52	**801.59**
4*20	5	1050.71	**1050.37**
	10	1010.34	**1001.79**
	15	996.09	**972.96**

References

Bonnabel, S. (2013).
Stochastic gradient descent on riemannian manifolds.
IEEE Transactions on Automatic Control, 58:2217–2229.

Nickel, M. and Kiela, D. (2017).
Poincaré embeddings for learning hierarchical representations.
In *Advances in neural information processing systems*, pages 6338–6347.

Wilson, B. and Leimeister, M. (2018).
Gradient descent in hyperbolic space.
arXiv preprint arXiv:1805.08207.

Zhang, H. and Sra, S. (2018).
Towards riemannian accelerated gradient methods.