# Hyperbolic Optimizer as a Dynamical System

**Nico Alvarado** [1 2 3] **Hans Lobel** [1 2 3 4]

## Abstract

During the last few years, the field of dynamical systems has been developing innovative tools to study the asymptotic behavior of different optimizers in the context of neural networks. In this work, we redefine an extensively studied optimizer, employing classical techniques from hyperbolic geometry. This new definition is linked to a non-linear differential equation as a continuous limit. Additionally, by utilizing Lyapunov stability concepts, we analyze the asymptotic behavior of its critical points.

## 1. Introduction

Classical optimization algorithms like gradient descent are commonly used, and their connection to dynamical systems is evident when viewing the weight updates as the evolution of a system state over iterations (Narendra & Parthasarathy, 1991). If $\theta$ represents the parameters, and $L(\theta)$ the loss function, the weight updates in each iteration, $\theta_{t+1} = \theta_t - \eta \nabla L(\theta_t)$, resemble the dynamics of a discrete-time dynamical system, where $\eta$ is the learning rate and $\nabla L(\theta_t)$ is the gradient of the loss function. The optimization process aims to locate minima within the loss landscape, analogous to identifying equilibrium points in the energy landscape of a dynamical system. This can be expressed as $\theta^* = \arg\min_\theta L(\theta)$, where $\theta^*$ signifies the optimal parameters.

Recall that hyperbolic spaces are homogeneous spaces of constant curvature equal to $-1$. The more relevant and crucial theoretical property of hyperbolic spaces and of spaces of negative curvature (Bridson & Haefliger, 2013) in general is that they can embed graphs such as trees with arbitrarily low distortion of the natural metrics. Gromov has first observed this (Gromov, 1987a), who introduced a much larger class of spaces, called $\delta$-hyperbolic spaces, which are shown to be almost isometric to trees (Gromov, 1987b), including cases of graphs with control on the diameter of cycles (Sarkar, 2012). In contrast, euclidean and positively curved spaces do not allow to embed trees with bounded distortion of the metric (Bourgain, 1985; Indyk et al., 2017; Chami, 2021; Aggarwal et al., 2001; Rodríguez-Flores & Papadopoulos, 2020; Borassi et al., 2015). Hyperbolic embeddings have also shown promise for routing (Cvetkovski & Crovella, 2009), clustering (Chami et al., 2020; Lamping et al., 1995), biological networks (Albert et al., 2014), phylogenetic trees (Billera et al., 2001; Matsumoto et al., 2021), neuroscience (Allard & Serrano, 2020), text embedding (Dhingra et al., 2018; Balazevic et al., 2019), knowledge graphs (Sun et al., 2020).

Hyperbolic Neural Networks (HNN) leverage hyperbolic geometry for representing data in a more natural way, especially for capturing hierarchical relationships (Chami et al., 2019; Ganea et al., 2018; Yang et al., 2022). However, the non-Euclidean nature of hyperbolic spaces poses challenges for classical optimizers. In the hyperbolic setting, if $\mathbf{g}(\theta_t)$ represent the Riemannian gradient, accounting for the curvature of the hyperbolic space, then the update rule becomes $\theta_{t+1} = \mathrm{Exp}_{\theta_t}(-\eta \mathbf{g}(\theta_t))$, where $\mathrm{Exp}_{\theta_t}$ is the exponential map in the hyperbolic space.

Optimizers tailored for hyperbolic geometry, such as Riemannian optimization methods, play a crucial role. They efficiently navigate the unique curvature of the hyperbolic space, ensuring stable convergence. Proper optimization allows hyperbolic neural networks to exploit their intrinsic geometry fully, leading to enhanced performance in capturing hierarchical relationships and complex data structures.

In this work, we present an optimizer based on ADMM (Boyd et al., 2011), but tailored to work in hyperbolic geometry, particularly within the Poincaré ball model. Establishing a connection between this optimizer and a non-linear ordinary differential equation (ODE) enriches our comprehension of the dynamics. The novel contribution lies in delving into stability through ODE linearization, offering valuable insights for practically implementing the hyperbolic optimizer in real-world applications.

[1]Department of Computer Science, Pontificia Universidad Católica de Chile [2]National Center of Artificial Intelligence, Chile [3]Millenium Institute Foundational Research on Data, Chile [4]Department of Transport and Logistics Engineering, Pontificia Universidad Católica de Chile. Correspondence to: Nico Alvarado <nfalvarado.mat@uc.cl>, Hans Lobel <halobel@uc.cl>.

*Proceedings of the 41st International Conference on Machine Learning*, Vienna, Austria. PMLR 235, 2024. 

### 1.1. Related work

The Alternating Direction Method of Multipliers (ADMM) can be seen as a dynamic process, whether we consider it as a continuous-time or discrete-time evolution. In the continuous case, ADMM updates are likened to the gradual transformation of states in a dynamical system ci. The algorithm employs an augmented Lagrangian function, and the iterative updates of Lagrange multipliers resemble the dynamics of continuous-time systems. Analyzing the continuous limit involves studying the algorithm's behavior through differential equations, shedding light on stability and convergence properties.

In practice, ADMM is often implemented as a discrete-time algorithm. Each iteration corresponds to a step in the evolution of a discrete-time dynamical system. This discrete nature allows us to understand ADMM through the framework of difference equations, capturing the recursive relationship between consecutive updates. The convergence of ADMM, akin to stability in dynamical systems, is often analyzed to provide assurances about the algorithm's reliability (Boyd et al., 2011). The fixed points or equilibria of ADMM correspond to solutions of the optimization problem, and understanding these points is analogous to analyzing stable states in a dynamical system.

The method often converges well for a wide range of convex optimization problems (França et al., 2018).

The scaled form of ADMM is given by (Boyd et al., 2011)

$$\begin{aligned} x_{k+1} &= \arg\min_{x\in\mathbb{R}^n} f(x) + \frac{\rho}{2}\|Ax - z_k + u_k\|^2 \\ z_{k+1} &= \arg\min_{z\in\mathbb{R}^m} g(z) + \frac{\rho}{2}\|Ax_{k+1} - z + u_k\|^2 \\ u_{k+1} &= u_k + Ax_{k+1} - z_{k+1}, \end{aligned}$$

where $\rho > 0$ is a penalty parameter and $u_k \in \mathbb{R}^m$ is the $k$th Lagrange multiplier estimate for the constrain $z = Ax$.

Let $f\colon \mathbb{R}^n \to \mathbb{R}$ and $g\colon \mathbb{R}^m \to \mathbb{R}$ be a continuously differentiable convex functions and $A \in \mathbb{R}^{m\times n}$ an invertible matrix.

**Theorem 1.1.** *(França et al., 2018) Consider the optimization problem given by*

$$\min_{x,z}\{V(x,z) = f(x) + g(z) \ \textit{subject to}\ z = Ax\}$$

*and the associated function $V(x) = f(x) + g(Ax)$. Then, the continuous limit associated with the ADMM updates, with time scale $t = k/\rho$, corresponds to the initial value problem*

$$X' + (A^TA)^{-1}\nabla V(X) = 0$$

*with $X(0) = x_0$.*

### 1.2. Paper contributions

Empirical evidence widely supports the effectiveness of low-dimensional hyperbolic spaces in learning hierarchical datasets. Despite a longstanding historical connection to embedding theory, as far as our knowledge extends, no theoretical investigations have been conducted on dynamical systems and hyperbolic optimizers. This article addresses and fills this gap in the literature, presenting the following key contributions:

- We proved the existence of a non-linear differential equation linked to the continuous limit of the Hyperbolic ADMM flow. This enables us to explore the asymptotic behavior of critical points and provides insights for conducting numerical analyses in future studies.
- We also proved that if a specific critical point $X^*$ remains at a low value under small perturbations, it signifies stability over time. This is advantageous as it indicates the optimization process is effective, steadily converging toward the best solution. The result offers a form of assurance that our optimization will ultimately settle at this optimal point and not deviate elsewhere.

## 2. Preliminaries

**Riemannian manifolds basics (see also (Petersen, 2016)).** We recall that a $d$-dimensional manifold $X$ is roughly a topological space that is locally parameterized by open sets of $\mathbb{R}^d$. A differentiable manifold has parametrizations such that the change of parametrization is a differentiable map. This allows us to define infinitesimal directions at each point $p \in X$, forming the tangent space $T_pX$ of $X$ at $p$. A differentiable manifold $X$ with an inner product $g_p(\cdot,\cdot)$ on each tangent space $T_pX \simeq \mathbb{R}^d$ (called a Riemannian metric) is a Riemannian manifold. By integrating the $g_{\gamma(t)}$-norm of the tangent vectors along a curve $\gamma(t)$, we can define the Riemannian length of a curve and the minimum length required to connect two points gives a Riemannian distance on $X$. A geodesic is a curve on a Riemannian manifold that locally minimizes the length between its endpoints.

The sectional curvature $k$ of a Riemannian manifold at a point $x$ in the tangent space $T_xM$ in the direction of two linearly independent tangent vectors $x, y$ is given by:

$$k(x,y) = \frac{\langle R(x,y)y, x\rangle}{\|x\|^2\|y\|^2 - \langle x,y\rangle^2},$$

where $R(x,y)y$ is the Riemann curvature tensor.

**Hyperbolic spaces.** Unlike Euclidean geometry, hyperbolic geometry rejects the parallel postulate, which leads to intriguing geometric properties. To fully understand this

geometry, one must become familiar with the hyperbolic parallel postulate and the concept of curvature. Furthermore, understanding how hyperbolic space is represented and visualized using models, such as the Poincaré disk or the hyperboloid model, is crucial.

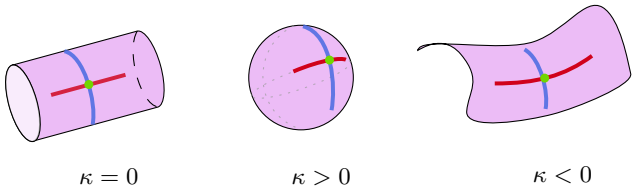


*Figure 1.* Example of different curvatures. On the right hand, we have negative curvature, thus a hyperbolic manifold.

The only Riemannian manifold of constant negative curvature $-1$ and dimension $d$ is the hyperbolic space $\mathbb{B}^d$, which can be identified (in the so-called Poincaré model) with the unit ball of $\mathbb{R}^d$ with the non-euclidean distance:

$$\gamma(x,y) = \operatorname{arccosh}\left(1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right).$$

Note that this distance is the Riemannian distance on the unit ball, associated with the Riemannian metric for which the norm of an infinitesimal vector $v$ at point $x$ is given by $\|v\|_x^2 = 2\|v\|/(1-\|x\|^2)$.

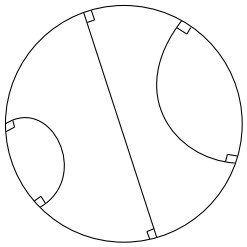

*Figure 2.* Poincaré model with $d = 2$. The curves are geodesics (i.e., coincide with the minimum-$\gamma$-length paths between any two points on the curve) and have infinite $\gamma$-length.

**Hyperbolic convexity.** In hyperbolic geometry, the concept of hyperbolic convexity plays a crucial role in understanding the properties of sets and functions. A set is hyperbolically convex if, for any two points within the set, the geodesic connecting them lies entirely within the set. Geodesics in hyperbolic space are represented by hyperbolic lines, the analogs of straight lines in Euclidean geometry.

Hyperbolically convex sets have unique properties, such as stability under hyperbolic isometries. This means that if a set is hyperbolically convex and undergoes a hyperbolic isometry, the transformed set remains hyperbolically convex. This property is essential in various applications, including understanding hyperbolic reflections and symmetry.

Moreover, the notion of hyperbolic convexity extends to functions. A function is considered hyperbolically convex if the region below its graph is a hyperbolically convex set. Hyperbolically convex functions have significant implications in optimization problems and variational principles in hyperbolic geometry.

The study of hyperbolically convex sets and functions provides valuable insights into the geometry and structure of hyperbolic space. It allows us to explore the relationships between geometric objects and transformations, offering a deeper understanding of the fundamental principles underlying hyperbolic geometry and its applications in diverse fields.

Convex sets in hyperbolic spaces, $\mathbb{H}^n$, are closely related to convex cones belonging to the interior of the Lorentz cone

$$\mathcal{L} := \left\{ x \in \mathbb{R}^{n+1} \colon x_{n+1} \geq \sqrt{x_1^2 + \cdots + x_n^2} \right\}.$$

**Definition 2.1.** We say that the set $\mathcal{C} \subseteq \mathbb{H}^n$ is hyperbolically convex if for any $p, q \in \mathcal{C}$ the geodesic segment joining $p$ to $q$ is contained in $\mathcal{C}$.

The hyperbolically convex sets are intersections of the hyperboloid with convex cones that belong to the interior of $\mathcal{L}$.

**Proposition 2.2.** *Let $\mathcal{C}$ be an open hyperbolically convex set and $f \colon \mathcal{C} \to \mathbb{R}$ be a differentiable function. The function $f$ is hyperbolically convex if and only if $f(q) \geq f(p) + \langle \nabla f(p), \log_p q \rangle$, for all $p, q \in \mathcal{C}$ and $p \neq q$.*

**Proposition 2.3.** *Let $\mathcal{C} \subseteq \mathbb{H}^n$ be an open hyperbolically convex set and $f \colon \mathcal{C} \to \mathbb{R}$ a differentiable function. The function $f$ is hyperbolically convex if and only if $\nabla f$ satisfies*

$$\langle \nabla f(p), \log_p q \rangle + \langle \nabla f(q), \log_q p \rangle \leq 0, \quad \forall p, q \in \mathcal{C},\ p \neq q.$$

**Gyrovector spaces.**

**Definition 2.4.** Let $V$ be a real inner product space and $V_s$ the ball centered in 0, of radius $s$. We define the Möbius addition as

$$u \oplus_s v = \frac{(1 + 2/s^2 u \cdot v + 1/s^2\|v\|^2)u + (1 - 1/s^2\|u\|^2)v}{1 + 2/s^2 u \cdot v + 1/s^4\|u\|^2\|v\|^2}.$$

Also we can define the Möbius subtraction as $u \ominus v := u \oplus (-v)$.

*Remark* 2.5. Note that if $s \to \infty$, then we can recover the Euclidean vector space sum.

In this work, we fix $s = 1$.

**Definition 2.6.** We define the Möbius scalar multiplication on $\mathbb{B}^n \setminus \{0\}$ by

$$\begin{aligned} r \otimes x &= \tanh(r\operatorname{arctanh}(\|x\|))\frac{x}{\|x\|} \\ &= \tanh\left(\frac{r}{2}\ln\left(\frac{1+\|x\|}{1-\|x\|}\right)\right)\frac{x}{\|x\|}. \end{aligned}$$

We say that a set of gyrovectors $\{\alpha_i\}_{i=1}^l$ is linearly independent in a gyrovector space if $r_1 \otimes \alpha_1 \oplus \cdots \oplus r_l \otimes \alpha_l = 0$ when $r_1 = r_2 = \cdots = r_l = 0$.

Given a matrix $M$ and a vector $x$ we define

$$M \otimes x = \tanh\left(\frac{\|Mx\|}{\|x\|}\operatorname{arctanh}(\|x\|)\right)\frac{Mx}{\|Mx\|}.$$

See Appendix A for more details.

## 3. Hyperbolic Optimization

### 3.1. Hyperbolic convex functions

We can extend the notion of hyperbolically convex functions to any hyperbolic model. Specifically, in the Poincaré ball model, we define

$$\begin{aligned} \log_x\colon \mathbb{B}^n &\to \mathcal{T}_x\mathbb{B}^n \\ y &\mapsto (1-\|x\|^2)\ln\left(\frac{1+\|-x\oplus y\|}{1-\|-x\oplus y\|}\right) \\ &\quad \cdot \frac{-x\oplus y}{\|-x\oplus y\|} \end{aligned}$$

and

$$\begin{aligned} \exp_x\colon \mathcal{T}_x\mathbb{B}^n &\to \mathbb{B}^n \\ y &\mapsto x \oplus \left(\tanh\left(\frac{\|y\|}{1-\|x\|^2}\right)\frac{y}{\|y\|}\right) \end{aligned}$$

If we take the tangent space in $x = 0$, we have

$$\begin{aligned} \log_0(x) &= \ln\left(\frac{1+\|x\|}{1-\|x\|}\right)\frac{x}{\|x\|} \\ \exp_0(x) &= \tanh\left(\|x\|\right)\frac{x}{\|x\|} \\ &= \frac{e^{2\|x\|}-1}{e^{2\|x\|}+1}\cdot\frac{x}{\|x\|}. \end{aligned}$$

**Definition 3.1.** A subset $S$ of a Riemannian manifold $\mathcal{M}$ is geodesically convex if, for every $x, y \in S$, there exists a geodesic segment $c\colon [0,1] \to \mathcal{M}$ such that $c(0) = x$, $c(1) = y$ and $c(t)$ is in $S$ for all $t \in [0,1]$.

In a geodesically convex set $S$, any two points are connected in $S$ by at least one geodesic segment $c$. Composing a function $f\colon S \to \mathbb{R}$ with $c$ yields a real function on $[0,1]$. If all of these compositions are convex in the usual sense, we say $f$ is convex in a geometric sense.

**Definition 3.2.** A function $f\colon S \to \mathbb{R}$ is geodesically (strictly) convex if $S$ is geodesically convex and $f \circ c\colon [0,1] \to \mathbb{R}$ is (strictly) convex for each geodesic segment $c\colon [0,1] \to \mathcal{M}$ whose image is in $S$ (with $c(0) \neq c(1)$).

In other words, for $S$ a geodesically convex set, we say $f\colon S \to \mathbb{R}$ is geodesically convex if for all $x, y \in S$ and all geodesics $c$ connecting $x$ to $y$ in $S$ the function $f \circ c\colon [0,1] \to \mathbb{R}$ is convex, that is,

$$\forall t \in [0,1], \quad f(c(t)) \leq (1-t)f(x) + tf(y).$$

It can be shown that if $g_x \in \mathcal{T}_x\mathcal{M}$ we have an equivalent definition of geodesically convex function:

$$f(x) \geq f(x) + \langle g_x + \exp_x^{-1}(y)\rangle_x, \quad \forall x, y \in \mathcal{M}.$$

If $\mathcal{M}$ is a hyperbolic manifold, we have

$$f(y) \geq f(x) + \langle \nabla f(x), \log_x y\rangle, \quad \forall x, y \in \mathcal{M}.$$

If $f$ satisfies the previous condition and $\mathcal{M}$ is a hyperbolic manifold, we say that $f$ is a hyperbolic convex function.

**Definition 3.3.** A function $f\colon \mathcal{M} \to \mathbb{R}$ is geodesically (hyperbolically) $\mu$-strongly convex if for any $x, y \in \mathcal{M}$,

$$f(y) \geq f(x) + \langle g_x, \exp_x^{-1} y\rangle_x + \frac{\mu}{2}d^2(x,y),$$

where $d(\cdot,\cdot)$ is the Riemannian distance.

If $\mathcal{M} = \mathbb{B}^n$ we have

$$\begin{aligned} f(y) &\geq f(x) + \langle \nabla f(x), \log_x y\rangle \\ &\quad + \frac{\mu}{2}\left(\operatorname{arccosh}\left(1 + \frac{2\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}\right)\right)^2, \end{aligned}$$

for any $x, y \in \mathcal{M}$.

It is clear that if $f$ is a hyperbolically $\mu$-strongly convex function, then it is hyperbolically strictly convex.

**Proposition 3.4.** *Let*

$$\begin{cases} \min f(x) \\ \textit{s.t. } x \in \mathbb{B}^n, \end{cases}$$

*where $f$ is hyperbolically convex. If $\eta \in \mathbb{B}^n$ satisfies $\nabla f(\eta) = 0$, then $\eta$ is a global minimum,*

*Proof.* From the definition of being hyperbolically convex, we have

$$f(y) \geq f(x) + \langle \nabla f(x), \log_x y\rangle, \qquad \forall x, y \in \mathbb{B}^n.$$

In particular, if we choose $x = \eta$, then

$$\begin{aligned} &f(y) \geq f(x) + \langle \nabla f(x), \log_x y \rangle, \qquad \forall x, y \in \mathbb{B}^n \\ &f(y) \geq f(\eta) + \langle \nabla f(\eta), \log_\eta y \rangle, \qquad \forall y \in \mathbb{B}^n. \\ &f(y) \geq f(\eta), \qquad \forall y \in \mathbb{B}^n. \end{aligned}$$

□

A necessary and sufficient condition for $\eta$ to be a global minimum is that $\nabla f(\eta) = 0$.

**Proposition 3.5.** *Let*

$$\begin{cases} \min f(x) \\ \textit{s.t. } x \in \Omega, \end{cases}$$

*be an optimization problem where $f \colon \Omega \subset \mathbb{B}^n \to \mathbb{R}$ is hyperbolically strictly convex on $\Omega$, a convex set. Then, the optimal solution is unique.*

*Proof.* Suppose that $x, y \in \mathbb{B}^n$ are different and both are solutions to the optimization problem. Then, $f(x) = f(y) \leq f(z)$, for any $z \in \Omega$. But since

$$f(y) > f(x) + \langle \nabla f(x), \log_x y \rangle,$$

and $\nabla f(x) = 0$ we have a contradiction. So, the solution must be unique. □

### 3.2. Hyperbolic ADMM

ADMM is an algorithm intended to blend the decomposability of dual ascent with the convergence properties of the method of multipliers. The algorithm solves problems in the form:

$$\begin{cases} \min f(x) + g(z) \\ \text{s.t. } Ax + Bz = c, \end{cases}$$

Let $f \colon \mathbb{B}^n \to \mathbb{R}$ and $g \colon \mathbb{B}^m \to \mathbb{R}$ be two continuously differentiable and hyperbolically convex functions. Now, choose a matrix $A \in \mathbb{B}^{m \times n}$ with full column rank, i.e., the columns vectors form a linearly independent set. Let $\oplus$ and $\otimes$ be the Möbius addition and multiplication defined in the Poincaré ball model.

Consider a function $V$ defined by:

$$\begin{aligned} V \colon \mathbb{B}^n &\to \mathbb{R} \\ x &\mapsto f(x) + g(A \otimes x). \end{aligned}$$

The following equations are the scaled form of ADMM in the gyrovector space version:

$$x_{k+1} = \underset{x \in \mathbb{B}^n}{\arg\min}\, f(x) + \frac{\rho}{2} \| A \otimes x \ominus z_k \oplus u_k \|^2 \quad (1)$$

$$z_{k+1} = \underset{z \in \mathbb{B}^m}{\arg\min}\, g(z) + \frac{\rho}{2} \| A \otimes x_{k+1} \ominus z \oplus u_k \|^2 \quad (2)$$

$$u_{k+1} = u_k \oplus A \otimes x_{k+1} \ominus z_{k+1}. \quad (3)$$

Unfortunately, the operations within gyrovector spaces present a challenging task due to their intricate nature. The inherent complexities make handling these spaces demanding and require a thoughtful approach. Due to this complexity, a better way to study ADMM in a hyperbolic space is by using a classical technique that identifies the Euclidean structure with the hyperbolic one.

Let $f \colon \mathbb{B}^n \to \mathbb{R}$ and $g \colon \mathbb{B}^m \to \mathbb{R}$ be two continuously differentiable and hyperbolically convex functions. Now, choose a matrix $A \in \mathbb{R}^{(m-1)\times(n-1)}$ with full column rank, i.e., the columns vectors form a linearly independent set.

$$\begin{aligned} V \colon \mathbb{B}^n &\to \mathbb{R} \\ x &\mapsto f(x) + g(\exp_y(A(\log_y x))). \end{aligned}$$

We define the scaled form of ADMM in the hyperbolic version:

$$x_{k+1} = \underset{x \in \mathbb{B}^n}{\arg\min}\, f(x) + \frac{\rho}{2} \| (A_y^{\otimes} x \ominus z_k) \oplus u_k \|^2 \quad (4)$$

$$z_{k+1} = \underset{z \in \mathbb{B}^m}{\arg\min}\, g(z) + \frac{\rho}{2} \| (A_y^{\otimes} x_{k+1} \ominus z) \oplus u_k \|^2 \quad (5)$$

$$u_{k+1} = u_k \oplus (A_y^{\otimes} x_{k+1}) \ominus z_{k+1}), \quad (6)$$

where $A_y^{\otimes} x = \exp_y(A(\log_y x))$.

In Riemannian Geometry, the choice of the tangent space at a particular point, often taken to be the zero point, is a convention that simplifies many calculations and allows for a more intuitive geometric interpretation. So,

$$x_{k+1} = \underset{x \in \mathbb{B}^n}{\arg\min}\, f(x) + \frac{\rho}{2} \| (A_0^{\otimes} x) \ominus z_k) \oplus u_k \|^2 \quad (7)$$

$$z_{k+1} = \underset{z \in \mathbb{B}^m}{\arg\min}\, g(z) + \frac{\rho}{2} \| (A_0^{\otimes} x_{k+1}) \ominus z) \oplus u_k \|^2 \quad (8)$$

$$u_{k+1} = u_k \oplus (A_0^{\otimes} x_{k+1}) \ominus z_{k+1}). \quad (9)$$

We have the main result of the paper:

**Theorem 3.6.** *Consider the hyperbolic optimization problem given by*

$$\begin{cases} \min_{x,z} \{ V(x, z) = f(x) + g(z) \}, \\ \textit{subject to } z = \exp_0(A(\log_0 x)) \end{cases}$$

*and the function $V(x) = f(x) + g(\exp_0(A(\log_0 x)))$. The continuous limit associated with the Hyperbolic ADMM*

*updates, with $t = k/\rho$, corresponds to the initial value problem*

$$(A^TA)^{-1}\nabla V(X) + (A^TA)^{-1}\Omega_1 + (A^TA)^{-1}\Omega_2 X + ((A^TA)^{-1}A\Omega_3 + \Omega_4 X)X' = 0 \tag{10}$$

*with $X(0) = x_0$*

*where $\Omega_i$ depends implicitly on $X$ for all $i = 1, 2, 3, 4$.*

*Sketch of the proof.* Define $L_\rho$ as

$$\begin{aligned} L_\rho \colon \mathbb{B}^n \times \mathbb{B}^m \times \mathbb{B}^m &\to \mathbb{R} \\ (x, z, u) &\mapsto f(x) + g(z) \\ &\quad + u^t(\exp_0(A(\log_0 x)) \ominus z) \\ &\quad + \frac{\rho}{2}\| \exp_0(A(\log_0 x)) \ominus z\|^2 \end{aligned}$$

Now, from Proposition 3.5, we have that if $(x_{k+1}, z_{k+1}, u_{k+1})$ satisfies equations 7, 8 and 9, then the solution is unique and then

$$\begin{aligned} 0 = \ell_{z_{k+1}}\nabla f(x_{k+1}) - \ell_{x_{k+1}}\nabla g(z_{k+1}) \\ + \rho\ell_{x_{k+1}}\ell_{z_{k+1}}\nu_{k+1}(z_{k+1} - z_k), \end{aligned}$$

for certain terms $\ell_{z_{k+1}}, \ell_{x_{k+1}}$ and $\nu_{k+1}$.

Following the idea of (França et al., 2018) we choose $t = \delta k$, $x_k = X(t)$, $z_k = Z(t)$, $u_k = U(t)$ and $\nu_k = N(t)$. Using the Mean Value Theorem on the $i$th component of $z_{k+1}$ we have that

$$z_{k+1,i} = Z_i(t + \delta) = Z_i(t) + \delta Z_i'(t + \zeta_i\delta),$$

for some $\zeta_i \in [0, 1]$. Thus

$$\lim_{\delta\to 0} \frac{z_{k+1,i} - z_{k,i}}{\delta} = Z_i'(t).$$

Since this hold for every $i$, we can choose $\rho = 1/\delta$ and get

$$\begin{aligned} 0 = \ell_{Z(t)}\nabla f(X(t)) - \ell_{X(t)}\nabla g(Z(t)) \\ + \ell_{X(t)}\ell_{Z(t)}N(t)Z'(t) \end{aligned}$$

Under the same idea, we can get $(\exp_0 A \log_0 X)_i(t) = Z_i(t)$ and since this holds for every component $i$ we have:

$$\begin{aligned} Z(t) &= \exp(A_0(\log_0 X(t))) \\ Z'(t) &= \eta(t) + \iota(t)X(t) + \kappa(t)AX'(t) \\ &\quad + \omega(t)A^TAX(t)X'(t), \end{aligned}$$

for certain terms $\eta(t), \iota(t), \kappa(t)$ and $\omega(t)$.

Finally, since nor $\ell_{X(t)}$ or $\ell_{Z(t)}$ vanishes at any point, we have the result. □

For more proof details, see Appendix B.

This theorem provides a way to understand the continuous version of the optimization process. In other words, it helps us predict how our variables will evolve smoothly over time as we try to find the best values to minimize the function $V(x, z)$. The initial value $X(0) = x_0$ gives us the starting point for this process.

## 4. Stability

Recall that $\gamma$ is the metric given in $\mathbb{B}^n$. Note that $(\mathbb{B}^n, \gamma)$ is a complete metric space. Fix $y \in \mathbb{B}^n$ and then,

$$\begin{aligned} \lim_{\|x\|\to 1} \gamma(x, y) &= \lim_{\tau\to\infty} \ln(\tau + \sqrt{\tau^2 - 1}) \\ &= \infty, \end{aligned}$$

where $\tau = 1 + 2\frac{\|x-y\|^2}{(1-\|x\|^2)(1-\|y\|^2)}$.

This implies that any Cauchy sequence is included in a set

$$K_r := \{r \colon \|y\| \le r < 1\}.$$

Clearly, $K_r$ is compact in the Euclidean topology. Also, the metrics are equivalent in $K_r$. Then we have convergence in the Euclidean sense if, and only if, we converge in the hyperbolic sense. Thus, $(\mathbb{B}^n, \gamma)$ is a complete metric space.

Now consider

$$X' = F(X, t), \quad X(t_0) = X_0 \tag{11}$$

a first order dynamical system with $F \colon \mathbb{B}^n \times \mathbb{R} \to \mathbb{B}^n$, $X = X(t) \in \mathbb{B}^n$, $X_0 \in \mathbb{B}^n$ and $t \in \mathbb{R}$. .

Let $F$ be a $L$-Lipschitz continuous function on $X$, i.e.

$$\gamma(F(X_1, t), F(X_2, t)) \le L\gamma(X_1, X_2)$$

for a fixed $t$, $L > 0$ and for all $X_1, X_2 \in \mathbb{B}^n$.

Let $\Omega \subseteq \mathbb{B}^n \times \mathbb{R}$, $(X_0, t_0) \in \Omega$ and suppose that $F$ is continuously differentiable on $\Omega$. Since $(\mathbb{B}^n, \gamma)$ is a complete metric space, 11 has a unique solution $X(t)$ on $t \in (t_0 - \varepsilon, t + \varepsilon)$ for any $\varepsilon > 0$ and $X(t_0) = X_0$. We can extend the solution throughout $\Omega$ and furthermore, due to (Hirsch et al., 2012). the solution is a continuous function of $(X_0, t_0)$, and if $F$ depends continuously on some set of parameters, then it's also a continuous function of those parameters.

**Definition 4.1.** Let $X^*$ be a point such that $F(X^*, t) = 0$ for all $t \ge t_0$. Then $X^*$ is a critical point of the system 11. Also:

1. The point $X^*$ is stable if for all neighborhood $\mathcal{O} \subseteq \mathbb{B}^n$ of $X^*$, there exists a neighborhood $\mathcal{O}' \subseteq \mathcal{O}$ of $X^*$

such that every solution $X(t)$ with initial condition $X(t_0) = X_0 \in \mathcal{O}'$ satisfies $F(\mathcal{O}', t) \subseteq \mathcal{O}$ for all $t > t_0$;

2. The point $X^*$ is asymptotically stable if is stable and, $\lim_{t\to\infty} X(t) = X^*$, for all $X_0 \in \mathcal{O}'$;

3. The point $X^*$ is unstable if it is not stable.

The given definition indicates that a critical point $X^*$ is considered stable if, within a small neighborhood, there exists an even smaller neighborhood $\mathcal{O}'$ where all solutions starting from points within $\mathcal{O}'$ remain within the original neighborhood $\mathcal{O}$ for all future times. If this stability condition also involves the system approaching $X^*$ as time goes to infinity, then it is termed asymptotically stable. Conversely, if $X^*$ is not stable, it is classified as unstable. This categorization provides insights into the long-term behavior and stability characteristics of the dynamic system centered around the critical point $X^*$.

The following result characterizes the critical points of the Hyperbolic ADMM flow 10.

**Proposition 4.2.** *Let $X^*$ be a strict local minimizer on $V(X)$. If $\|X^*(t)\| \to 1$ when $t \to \infty$, then $X^*$ is a critical point on the Hyperbolic ADMM flow 10.*

*Proof.* Since $X^*$ is a strict local minimizer for $V(X)$, then there exists a set $\mathcal{O}$ such that $X^* \in \mathcal{O}$ and $V(X) > V(X^*)$ for all $X \in \mathcal{O} \setminus \{X^*\}$. Due to the first-order optimality conditions, we have that $\nabla V(X^*) = 0$. Using this fact, and the fact that $\|X^*(t)\| \to 1$ when $t \to \infty$ we can conclude that $X^*$ is a critical point of the dynamical system 10. □

We can use the Lyapunov stability to check the stability of a system. In fact, we can determine if a dynamic system will stay in a particular state over time. Furthermore, we can extend to the Poincaré ball model a classical result (Hirsch et al., 2012).

**Theorem 4.3.** *(Hirsch et al., 2012) Let $X^*$ be a critical point of the dynamical system 11. Also, let $\mathcal{O} \subseteq \mathbb{B}^n$ be an open set containing $X^*$ and $\mathcal{L}\colon \mathcal{O} \to \mathbb{R}$ be a continuously differentiable function. We have the following:*

1. *if $\mathcal{L}(\cdot)$ satisfies*
   - *$\mathcal{L}(X^*) = 0$,*
   - *$\mathcal{L}(X) > 0$ for all $X \in \mathcal{O} \setminus \{X^*\}$,*
   - *$\mathcal{L}'(X) \leq 0$ for all $X \in \mathcal{O} \setminus \{X^*\}$,*

   *then $X^*$ is stable and $\mathcal{L}$ is called a Lyapunov function;*

2. *If we have a strict inequality in the last point, then $X^*$ is asymptotically stable, and $\mathcal{L}$ is called a strict Lyapunov function.*

The statement means that as a system evolves over time according to certain equations, a strict Lyapunov function can be used to show that the system's solutions decrease or get closer to a specific condition (see Figure 3). The level sets here refer to sets of points where the Lyapunov function takes constant values. The term strict implies that the Lyapunov function consistently decreases, emphasizing a clear trend towards stability in the system.

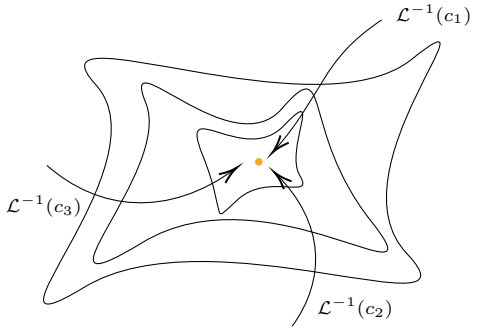


*Figure 3.* Solution decreases through the level sets of a strict Lyapunov function.

**Theorem 4.4.** *Let $X^*$ be a critical point (and a strict local minimizer of $V(X)$) of the linearized Hyperbolic ADMM flow. If $A$ is a positive definite matrix, $\nabla V(X) > 0$ near $X^*$ and $X(t) > 0$ for all $t > t_0$, where $t_0 \in [0, \infty)$, then $X^*$ is asymptotically stable.*

*Proof.* For this result, we need to assume that $n = m$, i.e., $A$ is a square matrix.

Recall that the flow of the Hyperbolic ADMM is given by

$$0 = (A^TA)^{-1}\nabla V(X) + (A^TA)^{-1}\Omega_1 + (A^TA)^{-1}\Omega_2 X + ((A^TA)^{-1}A\Omega_3 + \Omega_4 X)X'.$$

This is a non-linear differential equation. Then, if we take a small perturbation $X = X^* + \delta X$ implying $X' = \delta X'$ and replacing in the flow of the Hyperbolic ADMM, we have

$$0 = (A^TA)^{-1}\nabla V(X^* + \delta X) + (A^TA)^{-1}\Omega_1 + (A^TA)^{-1}\Omega_2(X^* + \delta X) + ((A^TA)^{-1}A\Omega_3 + \Omega_4(X^* + \delta X))\delta X'.$$

Now, if we linearize the non-linear terms by neglecting the

high-order terms involving $\delta X$ and $\delta X'$, we have

$$\begin{aligned} 0 &= (A^TA)^{-1}\nabla^2 V(X^*)\delta X + (A^TA)^{-1}\delta X \\ &\quad + ((A^TA)^{-1}A\Omega_3 + \Omega_4 X^*)\delta X'. \\ 0 &= \underbrace{(A^TA)^{-1}(\nabla^2 V(X^*) + \mathbb{I}_{n\times n})}_{=\tilde{A}}\delta X \\ &\quad + \underbrace{((A^TA)^{-1}A\Omega_3 + \Omega_4 X^*)}_{=\tilde{B}}\delta X'. \end{aligned}$$

Thus, we have a dynamical system of the form

$$\tilde{A}X + \tilde{B}X' = 0. \tag{12}$$

Note that in 12, we're omitting $\delta$.

Finally, defining $\mathcal{L}(X) = X^TPX$, where $P$ is a positive definite matrix, we can show using Theorem 4.3, and the fact that $A$ is a positive definite matrix, that the critical point $X^*$ is asymptotically stable. □

This result states that if we have a specific critical point $X^*$ and this point stays low when we make small changes around it, then it is stable over time.

In the context of the Hyperbolic ADMM flow, this means that if we start at $X^*$ and the conditions mentioned in the result are satisfied, then as time goes on, we will stay close to the critical point. This is good because it indicates that the optimization process works effectively, converging to the best solution. The result provides a kind of guarantee that our optimization will eventually settle at this optimal point and not wander away.

## 5. Conclusions and future work

In this study, we introduce a novel optimizer using hyperbolic geometry. Specifically, we connect the Poincaré ball model to a non-linear differential equation. The complexity arises when the equation is not linearized, necessitating numerical analysis for stability studies. Linearizing the ODE reveals crucial insights into system behavior.

Looking forward, we propose a more general exploration of the optimizer, incorporating exponential and logarithmic operations at arbitrary points. This broadens our understanding of its behavior. We also pose questions about extending the optimizer to other hyperbolic models and the impact of isometries on stability and convergence.

Comparing our hyperbolic ADMM with the original, we find increased complexity in the hyperbolic version, advancing our understanding of why Hyperbolic Neural Networks perform well. We anticipate a numerical comparison with ADMM in gyrovector space, suggesting our hyperbolic version's potential superiority. While Möbius operations are computationally expensive, their optimization may be task-dependent, offering a balance between cost and efficiency.

## Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

## Acknowledgements

The authors thank Sebastian Burgos for helpful comments and conversations.

The authors also acknowledge funding support from the Millennium Institute Foundational Research on Data, Chile (IMFD ANID - Millennium Science Initiative Program - Code ICN17_002) and the National Center of Artificial Intelligence, Chile (CENIA FB210017, Basal ANID).

## References

Aggarwal, C. C., Hinneburg, A., and Keim, D. A. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pp. 420–434. Springer, 2001.

Albert, R., DasGupta, B., and Mobasheri, N. Topological implications of negative curvature for biological and social networks. *Physical Review E*, 89(3):032811, 2014.

Allard, A. and Serrano, M. Á. Navigable maps of structural brain networks across species. *PLoS computational biology*, 16(2):e1007584, 2020.

Balazevic, I., Allen, C., and Hospedales, T. Multi-relational poincaré graph embeddings. In *Proceedings of NeurIPS*, pp. 4463–4473, 2019.

Billera, L. J., Holmes, S. P., and Vogtmann, K. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

Borassi, M., Chessa, A., and Caldarelli, G. Hyperbolicity measures democracy in real-world networks. *Physical Review E*, 92(3):032812, 2015.

Bourgain, J. On lipschitz embedding of finite metric spaces in hilbert space. *Israel Journal of Mathematics*, 52:46–52, 1985.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 1(1):1–122, 2011. doi: 10.1561/2200000016.

Bridson, M. R. and Haefliger, A. *Metric spaces of non-positive curvature*, volume 319. Springer Science & Business Media, 2013.

Chami, I. *Representation Learning and Algorithms in Hyperbolic Spaces*. Stanford University, 2021.

Chami, I., Ying, Z., Ré, C., and Leskovec, J. Hyperbolic Graph Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076, 2020.

Cvetkovski, A. and Crovella, M. Hyperbolic embedding and routing for dynamic graphs. In *IEEE INFOCOM 2009*, pp. 1647–1655. IEEE, 2009.

Dhingra, B., Shallue, C. J., Norouzi, M., Dai, A. M., and Dahl, G. E. Embedding text in hyperbolic spaces. *arXiv preprint arXiv:1806.04313*, 2018.

França, G., Robinson, D., and Vidal, R. Admm and accelerated admm as continuous dynamical systems. 05 2018.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Gromov, M. *Hyperbolic groups*. Springer, 1987a.

Gromov, M. Hyperbolic Groups. In Gersten, S. M. (ed.), *Essays in Group Theory*, Mathematical Sciences Research Institute Publications, pp. 75–263. Springer, New York, NY, 1987b. ISBN 978-1-4613-9586-7. doi: 10.1007/978-1-4613-9586-7_3. URL https://doi.org/10.1007/978-1-4613-9586-7_3.

Hirsch, M., Smale, S., and Devaney, R. *Differential Equations, Dynamical Systems, and an Introduction to Chaos*. Academic Press, 2012.

Indyk, P., Matou ˇsek, J., and Sidiropoulos, A. 8 LOW-DISTORTION EMBEDDINGS OF FINITE METRIC SPACES. 2017.

Lamping, J., Rao, R., and Pirolli, P. A focus+context technique based on hyperbolic geometry for visualizing large hierarchies. In *Proceedings of the SIGCHI conference on Human factors in computing systems - CHI '95*, pp. 401–408, Denver, Colorado, United States, 1995. ACM Press. ISBN 978-0-201-84705-5. doi: 10.1145/223904.223956. URL http://portal.acm.org/citation.cfm?doid=223904.223956.

Matsumoto, H., Mimori, T., and Fukunaga, T. Novel metric for hyperbolic phylogenetic tree embeddings. *Biology Methods and Protocols*, 6(1):bpab006, 2021.

Narendra, K. and Parthasarathy, K. Gradient methods for the optimization of dynamical systems containing neural networks. *IEEE Transactions on Neural Networks*, 2(2): 252–262, 1991. doi: 10.1109/72.80336.

Petersen, P. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer International Publishing, Cham, 2016. ISBN 978-3-319-26652-7 978-3-319-26654-1. doi: 10.1007/978-3-319-26654-1. URL http://link.springer.com/10.1007/978-3-319-26654-1.

Rodríguez-Flores, M. A. and Papadopoulos, F. Hyperbolic mapping of human proximity networks. *Scientific Reports*, 10(1):20244, 2020.

Sarkar, R. Low Distortion Delaunay Embedding of Trees in Hyperbolic Plane. In Van Kreveld, M. and Speckmann, B. (eds.), *Graph Drawing*, volume 7034, pp. 355–366. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012. ISBN 978-3-642-25877-0 978-3-642-25878-7. doi: 10.1007/978-3-642-25878-7_34. URL http://link.springer.com/10.1007/978-3-642-25878-7_34. Series Title: Lecture Notes in Computer Science.

Sun, Z., Chen, M., Hu, W., Wang, C., Dai, J., and Zhang, W. Knowledge association with hyperbolic knowledge graph embeddings. *arXiv preprint arXiv:2010.02162*, 2020.

Yang, M., Zhou, M., Li, Z., Liu, J., Pan, L., Xiong, H., and King, I. Hyperbolic graph neural networks: A review of methods and applications. *arXiv preprint arXiv:2202.13852*, 2022.

## A. Use of Möbius operations

**Definition A.1.** Two vectors $x, y \in \mathbb{B}^n$ are linearly dependent if for some $a \in \mathbb{R}$ we can write $x = a \otimes y$.

Consider the set $\{(1/2, 0), (1/4, 0)\}$. Now,

$$\begin{aligned}\begin{pmatrix}1/2\\0\end{pmatrix} &= a \otimes \begin{pmatrix}1/4\\0\end{pmatrix}\\ &= \tanh\left(\frac{a}{2}\ln\left(\frac{1+1/4}{1-1/4}\right)\right) 4\begin{pmatrix}1/4\\0\end{pmatrix}\\ &= \left(\frac{(5/3)^a - 1}{(5/3)^a + 1}\right)\begin{pmatrix}1\\0\end{pmatrix}\end{aligned}$$

Then we have

$$\begin{aligned}\frac{1}{2} &= \frac{(5/3)^a - 1}{(5/3)^a + 1}\\ (5/3)^a + 1 &= 2(5/3)^a - 2\\ a &= \frac{\ln 3}{\ln(5/3)}.\end{aligned}$$

Thus, the vectors $(1/2, 0), (1/4, 0)$ are linearly dependent.

More generally, if we have

$$\begin{pmatrix}x\\0\end{pmatrix}, \begin{pmatrix}y\\0\end{pmatrix} \in \mathbb{B}^2,$$

then

$$\begin{aligned}\begin{pmatrix}x\\0\end{pmatrix} &= a \otimes \begin{pmatrix}y\\0\end{pmatrix}\\ x &= \left(\frac{\left(\frac{1+|y|}{1-|y|}\right)^a - 1}{\left(\frac{1+|y|}{1-|y|}\right)^a + 1}\right)\frac{y}{|y|}\\ a &= \frac{\ln\frac{x+y/|y|}{y/|y|-x}}{\ln\frac{1+|y|}{1-|y|}} \in \mathbb{R}\end{aligned}$$

**Proposition A.2.** *A set $A$ is linearly dependent in $\mathbb{R}^n$ if and only if is linearly dependent in $\mathbb{B}^n$*

In $\mathbb{B}^n$, choose $\{(1/2, 0), (0, 1/2)\}$ and suppose that $a, b \neq 0$, then

$$\begin{aligned}a \otimes \begin{pmatrix}1/2\\0\end{pmatrix} \oplus b \otimes \begin{pmatrix}0\\1/2\end{pmatrix} &= 0\\ \left(\frac{2\cdot 3^{2b} + 2}{(3^b+1)^2}\right)\begin{pmatrix}1\\0\end{pmatrix} + \left(\frac{4\cdot 3^a}{(3^a+1)^2}\right)\begin{pmatrix}0\\1\end{pmatrix} &= 0\end{aligned}$$

Thus,

$$\begin{aligned}3^{2b} + 1 &= 0\\ 3^a &= 0.\end{aligned}$$

The previous system has no solutions. Thus, the set $\{(1/2,0),(0,1/2)\}$ is l.i.

Now, if we choose $\{(1/2,0),(1/2,0)\}$ and again $a,b\neq 0$ we will have

$$\begin{aligned}
1+2\left(\frac{3^a-1}{3^a+1}\right)\left(\frac{3^b-1}{3^b+1}\right)+\left(\frac{3^b-1}{3^b+1}\right)^2+1-\left(\frac{3^a-1}{3^a+1}\right)^2&=0\\
2-2\left(\frac{3^a-1}{3^a+1}\right)^2+\left(\left(\frac{3^a-1}{3^a+1}\right)+\left(\frac{3^b-1}{3^b+1}\right)\right)^2&=0\\
\left(\left(\frac{3^a-1}{3^a+1}\right)+\left(\frac{3^b-1}{3^b+1}\right)\right)^2&=2\left(\frac{3^a-1}{3^a+1}\right)^2-2\\
(x+y)^2&=2x^2-2,
\end{aligned}$$

where

$$x=\frac{3^a-1}{3^a+1},\quad y=\frac{3^b-1}{3^b+1}.$$

This equation has no solutions. Note that

$$\begin{aligned}
(x+y)^2&=2x^2-2\\
y&=\pm\sqrt{2}\sqrt{x^2-1}-x
\end{aligned}$$

and $x^2-1<0$. In fact, it is easy to note that

$$\begin{pmatrix}1/2\\0\end{pmatrix}=1\otimes\begin{pmatrix}1/2\\0\end{pmatrix}$$

**Proposition A.3.** *Let $\mathbb{B}^n=(\mathbb{B}^n,\oplus,\otimes)$. If a set of two vectors $x,y$ are orthogonal in $\mathbb{R}^n$, then $x,y$ are linearly independent in $\mathbb{B}^n$.*

*Proof.* Consider $a,b\in R\setminus\{0\}$ and $\{x,y\}\subseteq\mathbb{B}^n$ such that $\langle x,y\rangle=0$. Then,

$$\begin{aligned}
a\otimes x\oplus b\otimes y&=0\\
\tanh\left(\frac{a}{2}\ln\left(\frac{1+\|x\|}{1-\|x\|}\right)\right)\frac{x}{\|x\|}\oplus\tanh\left(\frac{b}{2}\ln\left(\frac{1+\|y\|}{1-\|y\|}\right)\right)\frac{y}{\|y\|}&=0\\
\underbrace{\left(1+\frac{\left(\left(\frac{1+\|y\|}{1-\|y\|}\right)^b-1\right)^2}{\left(\left(\frac{1+\|y\|}{1-\|y\|}\right)^b+1\right)^2}\right)}_{=r_1}x+\underbrace{\left(1+\frac{\left(\left(\frac{1+\|x\|}{1-\|x\|}\right)^a-1\right)^2}{\left(\left(\frac{1+\|x\|}{1-\|x\|}\right)^a+1\right)^2}\right)}_{=r_2}y&=0
\end{aligned}$$

It is easy to see that $r_1$ and $r_2$ can't be zero. So, the orthogonal set $\{x,y\}$ is linearly independent. □

For a matrix multiplication, we will use the following example.

Take

$$M=\begin{pmatrix}1/2&0\\0&1/2\end{pmatrix},\quad N=\begin{pmatrix}1/3&0\\0&1/3\end{pmatrix}.$$

Then

$$M\otimes N=\begin{pmatrix}\frac{\sqrt{2}-1}{\sqrt{2}+1}&0\\0&\frac{\sqrt{2}-1}{\sqrt{2}+1}\end{pmatrix}$$

If $M, N \in \mathbb{B}^{n\times m}$ where $M = [M_1|\dots|M_m]$ and $N = [N_1|\dots|N_m]$ we define

$$M \oplus N = [M_1 \oplus N_1|\dots|M_m \oplus N_m].$$

As an example, choose

$$M = \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad N = \begin{pmatrix} 1/3 & 0 \\ 0 & 1/3 \end{pmatrix}.$$

Then

$$\begin{aligned} M \oplus N &= \begin{pmatrix} 1/2 & 0 \\ 0 & 1/2 \end{pmatrix} \oplus \begin{pmatrix} 1/3 & 0 \\ 0 & 1/3 \end{pmatrix} \\ &= \begin{pmatrix} 2/3 & 0 \\ 0 & 2/3 \end{pmatrix} \end{aligned}$$

Let $x \in \mathbb{B}^n$. We're searching for a matrix $M$ such that

$$M \otimes x = x.$$

If we choose $M$ as the $n$-dimensional identity matrix (i.e. diagonal entries are 1s and the rest 0s) we have

$$\begin{aligned} M \otimes x &= x \\ \tanh\left(\ln\left(\frac{1+\|x\|}{1-\|x\|}\right)\right)\frac{x}{\|x\|} &= x \\ \left(\frac{2}{1+\|x\|^2}\right)x &= x \end{aligned}$$

then,

$$\begin{aligned} \frac{2x_1}{1+x_1^2+x_2^2} &= x_1 \\ \frac{2x_2}{1+x_1^2+x_2^2} &= x_2. \end{aligned}$$

This equation has no solutions (assuming that $x_1 \neq x_2 \neq 0$) because $\|x\| < 1$. Thus, $M$ cannot be the $n$-identity matrix.

To solve $M \otimes x = x$ for $M$, we have to compute when

$$\tanh\left(\frac{\|Mx\|}{\|x\|}\operatorname{arctanh}(\|x\|)\right) = \|Mx\|.$$

Then

$$2\frac{\|Mx\|}{\|x\|}\ln\left(\frac{1+\|x\|}{1-\|x\|}\right) = \ln\left(\frac{1+\|Mx\|}{1-\|Mx\|}\right).$$

This equation has solutions only by numerical approximation.

## B. Proof of Theorem 3.6

**Theorem B.1.** *Consider the hyperbolic optimization problem given by*

$$\begin{cases} \min_{x,z}\{V(x,z) = f(x) + g(z)\}, \\ \textit{subject to } z = \exp_0(A(\log_0 x)) \end{cases}$$

*and the function $V(x) = f(x) + g(\exp_0(A(\log_0 x)))$. The continuous limit associated with the Hyperbolic ADMM updates, with $t = k/\rho$, corresponds to the initial value problem*

$$(A^TA)^{-1}\nabla V(X) + (A^TA)^{-1}\Omega_1 + (A^TA)^{-1}\Omega_2 X + ((A^TA)^{-1}A\Omega_3 + \Omega_4 X)X' = 0 \tag{13}$$

*with $X(0) = x_0$*

*where $\Omega_i$ depends implicitly on $X$ for all $i = 1, 2, 3, 4$.*

*Proof.* Define $L_\rho$ as

$$\begin{aligned} L_\rho \colon \mathbb{B}^n \times \mathbb{B}^m \times \mathbb{B}^m &\to \mathbb{R} \\ (x, z, u) &\mapsto f(x) + g(z) + u^t(\exp_0(A(\log_0 x)) \ominus z) + \frac{\rho}{2}\| \exp_0(A(\log_0 x)) \ominus z\|^2 \end{aligned}$$

Differentiating $L_\rho$ w.r.t $u$ we have

$$\begin{aligned} \frac{\partial L_\rho}{\partial u} &= u_k \oplus (\exp_0(A(\log_0 x_{k+1})) \ominus z_{k+1}) \\ u_{k+1} &= u_k \oplus \left( \exp_0 A \left( \ln \left( \frac{1 + \|x_{k+1}\|}{1 - \|x_{k+1}\|} \right) \frac{x_{k+1}}{\|x_{k+1}\|} \right) \ominus z_{k+1} \right) \\ &= u_k \oplus \left( \exp_0 \ln \left( \frac{1 + \|x_{k+1}\|}{1 - \|x_{k+1}\|} \right) \frac{Ax_{k+1}}{\|x_{k+1}\|} \ominus z_{k+1} \right) \\ &= u_k \oplus \left( \frac{e^{2\ln\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} - 1}{e^{2\ln\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} + 1} \cdot \frac{Ax_{k+1}}{\|Ax_{k+1}\|} \ominus z_{k+1} \right) \\ &= u_k \oplus \left( \frac{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} - 1}{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} + 1} \cdot \frac{Ax_{k+1}}{\|Ax_{k+1}\|} \ominus z_{k+1} \right). \end{aligned}$$

Now define

$$\alpha_{k+1} = \frac{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} - 1}{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} + 1} \cdot \frac{Ax_{k+1}}{\|Ax_{k+1}\|}.$$

Then,

$$\begin{aligned} u_{k+1} &= u_k \oplus (\alpha_{k+1} \ominus z_{k+1}) \\ &= u_k \oplus \left( \frac{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)\alpha_{k+1} - (1 - \|\alpha_{k+1}\|^2)z_{k+1}}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2} \right) \\ &= u_k \oplus \left( \frac{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2}\alpha_{k+1} - \frac{1 - \|\alpha_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2} z_{k+1} \right). \end{aligned}$$

In the right hand of the equality, we have two constants multiplying two vectors, call it $\mu_{k+1}$ and $\nu_{k+1}$ respectively. Thus,

$$u_{k+1} = \frac{(1 + 2\langle u_k, \mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\rangle + \|\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\|^2)u_k + (1 - \|u_k\|^2)(\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1})}{1 - 2\langle u_k, \mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\rangle + \|u_k\|^2\|\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\|^2}.$$

Using the previous notation we can redefine equations 7 and 8 as:

$$x_{k+1} = \arg\min_{x \in \mathbb{B}^n} f(x) + \frac{\rho}{2} \|(\mu\alpha - \nu_k z_k) \oplus u_k\|^2 \tag{14}$$

$$z_{k+1} = \arg\min_{z \in \mathbb{B}^m} g(z) + \frac{\rho}{2} \|(\mu_{k+1}\alpha_{k+1} - \nu z) \oplus u_k\|^2 \tag{15}$$

Now, from Proposition 3.5, we have that if $(x_{k+1}, z_{k+1}, u_{k+1})$ satisfies equations 14, 15 and 9, then the solution is unique. Thus, we have

$$0 = \nabla f(x_{k+1}) + \rho(\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_k + u_{k+1}) \underbrace{\left(\frac{\partial \mu_{k+1}}{\partial x_{k+1}}\alpha_{k+1} + \frac{\partial \alpha_{k+1}}{\partial x_{k+1}}\mu_{k+1} - \frac{\partial \nu_{k+1}}{\partial x_{k+1}}z_{k+1}\right)}_{=\ell_{x_{k+1}}}$$

$$0 = \nabla g(z_{k+1}) + \rho(\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1} + u_{k+1}) \underbrace{\left(\frac{\partial \mu_{k+1}}{\partial z_{k+1}}\alpha_{k+1} - \frac{\partial \nu_{k+1}}{\partial z_{k+1}}z_{k+1} - \nu_{k+1}\right)}_{=\ell_{z_{k+1}}}$$

Multiplying the first equality by $\ell_{z_{k+1}}$ , the second one by $\ell_{x_{k+1}}$ and subtracting both we have

$$0 = \ell_{z_{k+1}} \nabla f(x_{k+1}) - \ell_{x_{k+1}} \nabla g(z_{k+1}) + \rho \ell_{x_{k+1}} \ell_{z_{k+1}} \nu_{k+1}(z_{k+1} - z_k).$$

Following the idea of (Boyd et al., 2011) we choose $t = \delta k$, $x_k = X(t)$, $z_k = Z(t)$, $u_k = U(t)$ and $\nu_k = N(t)$. Using the Mean Value Theorem on the $i$th component of $z_{k+1}$ we have that

$$z_{k+1,i} = Z_i(t + \delta) = Z_i(t) + \delta Z_i'(t + \zeta_i \delta), \quad \text{for some } \zeta_i \in [0, 1].$$

Thus

$$\lim_{\delta \to 0} \frac{z_{k+1,i} - z_{k,i}}{\delta} = Z_i'(t).$$

Since this hold for every $i$, we can choose $\rho = 1/\delta$ and get

$$\ell_{z_{k+1}} \nabla f(x_{k+1}) - \ell_{x_{k+1}} \nabla g(z_{k+1}) + \rho \ell_{x_{k+1}} \ell_{z_{k+1}} \nu_{k+1}(z_{k+1} - z_k) = 0$$
$$\rightarrow$$
$$\ell_{Z(t)} \nabla f(X(t)) - \ell_{X(t)} \nabla g(Z(t)) + \ell_{X(t)} \ell_{Z(t)} N(t) Z'(t) = 0$$

Now, on the $i-$th component of 9 we have

$$U_i(t + \delta) = U_i(t) + (\exp_0(A(\log_0 X)))_i(t + \delta) - Z_i(t + \delta).$$

Again, by the Mean Value Theorem there exists $\zeta_i \in [0, 1]$ such that

$$\delta U_i'(t + \zeta_i \delta) = (\exp_0 A \log_0 X)_i(t + \delta) - Z_i(t + \delta)$$

so, $(\exp_0 A \log_0 X)_i(t) = Z_i(t)$. Since this holds for every component $i$, we have:

$$
\begin{aligned}
Z(t) &= \exp_0(A(\log_0 X(t))) \\
Z'(t) &= -2\frac{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}\left(\log\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)2\frac{\|AX(t)\|}{\|X(t)\|}\frac{\|A\|}{\|X(t)\|}+\frac{2\frac{\|AX(t)\|}{\|X(t)\|}}{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)}\frac{-2X(t)}{(1-\|X(t)\|)^2}\right)}{\left(\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1\right)^2} \\
&+\left(\frac{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}+1}{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1}\right)\left(\frac{A\|AX(t)\|-A^TAX(t)}{\|AX(t)\|^2}\right)X'(t) \\
&= \eta(t)+\iota(t)X(t)+\kappa(t)AX'(t)+\omega(t)A^TAX(t)X'(t),
\end{aligned}
$$

where

$$
\begin{aligned}
\eta(t) &= -4\frac{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}\log\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)\frac{\lambda_{max}(A^TA)}{\|X(t)\|}}{\left(\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1\right)^2} \\
\iota(t) &= 8\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}\frac{\frac{\|AX(t)\|}{(1-\|X(t)\|^2)\|X(t)\|}}{\left(\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1\right)^2} \\
\kappa(t) &= \left(\frac{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}+1}{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1}\right)\frac{1}{\|AX(t)\|} \\
\omega(t) &= -\left(\frac{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}+1}{\left(\frac{1+\|X(t)\|}{1-\|X(t)\|}\right)^{2\frac{\|AX(t)\|}{\|X(t)\|}}-1}\right)\frac{1}{\|AX(t)\|^2}.
\end{aligned}
$$

Since nor $\ell_{X(t)}$ or $\ell_{Z(t)}$ vanishes at any point, we have

$$
\begin{aligned}
\nabla f(X(t)) - \frac{\ell_{X(t)}}{\ell_{Z(t)}}\nabla g(Z(t)) + \ell_{X(t)}N(t)Z'(t) &= 0 \\
\nabla V(X(t)) + \underbrace{\ell_{X(t)}N(t)\eta(t)}_{=\Omega_1(t)} + \underbrace{\ell_{X(t)}N(t)\iota(t)}_{=\Omega_2(t)}X(t) + \underbrace{\ell_{X(t)}N(t)\kappa(t)}_{=\Omega_3(t)}AX'(t) + \underbrace{\ell_{X(t)}N(t)\omega(t)}_{=\Omega_4(t)}A^TAX(t)X'(t)) &= 0 \\
\nabla V(X(t)) + \Omega_1(t) + \Omega_2(t)X(t) + \Omega_3(t)AX'(t) + \Omega_4(t)A^TAX(t)X'(t)) &= 0 \\
(A^TA)^{-1}\nabla V(X(t)) + (A^TA)^{-1}\Omega_1(t) + (A^TA)^{-1}\Omega_2(t)X(t) + (A^TA)^{-1}A\Omega_3(t)X'(t) + \Omega_4(t)X(t)X'(t) &= 0 \\
(A^TA)^{-1}\nabla V(X(t)) + (A^TA)^{-1}\Omega_1(t) + (A^TA)^{-1}\Omega_2(t)X(t) + ((A^TA)^{-1}A\Omega_3(t) + \Omega_4(t)X(t))X'(t) &= 0,
\end{aligned}
$$

which is equivalent to 10 since $A$ is invertible.

Finally, since 10 is a non-homogeneous first-order linear equation, the dynamics is specified by $X(0) = x_0$, where $x_0$ is the estimate of the initial solution of B.1. □

## C. Closed forms.

Due the Theorem 3.6 we have explicit forms for the derivatives of $\alpha_{k+1}$, $\mu_{k+1}$ and $\nu_{k+1}$ w.r.t. $x_{k+1}$ and $z_{k+1}$. This will be useful to run experiments in the future.

Recall

$$u_{k+1} - \frac{(1 + 2\langle u_k, \mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\rangle + \|\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\|^2)u_k + (1 - \|u_k\|^2)(\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1})}{1 - 2\langle u_k, \mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\rangle + \|u_k\|^2\|\mu_{k+1}\alpha_{k+1} - \nu_{k+1}z_{k+1}\|^2} = 0$$

For $\alpha_{k+1}$ define

$$\beta_{k+1} = \frac{1 + \|x_{k+1}\|}{1 - \|x_{k+1}\|}, \quad \gamma_{k+1} = 2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|} \text{ and } \delta_{k+1} = \frac{Ax_{k+1}}{\|Ax_{k+1}\|}$$

this implies that

$$\alpha_{k+1} = \left(\frac{\beta_{k+1}^{\gamma_{k+1}} + 1}{\beta_{k+1}^{\gamma_{k+1}} - 1}\right)\delta_{k+1}$$

Then,

$$\frac{\partial \beta_{k+1}}{\partial x_{k+1}} = \frac{-2x_{k+1}}{(1 - \|x_{k+1}\|)^2}, \quad \frac{\partial \gamma_{k+1}}{\partial x_{k+1}} = 2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}\frac{\|A\|}{\|x_{k+1}\|} \text{ and } \frac{\partial \delta_{k+1}}{\partial x_{k+1}} = \frac{A\|Ax_{k+1}\| - A^TAx_{k+1}}{\|Ax_{k+1}\|^2}$$

and

$$\begin{aligned}
\frac{\partial \alpha_{k+1}}{\partial x_{k+1}} &= \frac{\beta_{k+1}^{\gamma_{k+1}}\left(\log\beta_{k+1}\frac{\partial\gamma_{k+1}}{\partial x_{k+1}} + \frac{\gamma_{k+1}}{\beta_{k+1}}\frac{\partial\beta_{k+1}}{\partial x_{k+1}}\right)(\beta_{k+1}^{\gamma_{k+1}} - 1)}{(\beta_{k+1}^{\gamma_{k+1}} - 1)^2} \\
&\quad - \frac{\beta_{k+1}^{\gamma_{k+1}}\left(\log\beta_{k+1}\frac{\partial\gamma_{k+1}}{\partial x_{k+1}} + \frac{\gamma_{k+1}}{\beta_{k+1}}\frac{\partial\beta_{k+1}}{\partial x_{k+1}}\right)(\beta_{k+1}^{\gamma_{k+1}} + 1)}{(\beta_{k+1}^{\gamma_{k+1}} - 1)^2} \\
&\quad + \left(\frac{\beta_{k+1}^{\gamma_{k+1}} + 1}{\beta_{k+1}^{\gamma_{k+1}} - 1}\right)\frac{\partial\delta_{k+1}}{\partial x_{k+1}} \\
&= -2\frac{\beta_{k+1}^{\gamma_{k+1}}\left(\log\beta_{k+1}\frac{\partial\gamma_{k+1}}{\partial x_{k+1}} + \frac{\gamma_{k+1}}{\beta_{k+1}}\frac{\partial\beta_{k+1}}{\partial x_{k+1}}\right)}{(\beta_{k+1}^{\gamma_{k+1}} - 1)^2} + \left(\frac{\beta_{k+1}^{\gamma_{k+1}} + 1}{\beta_{k+1}^{\gamma_{k+1}} - 1}\right)\frac{\partial\delta_{k+1}}{\partial x_{k+1}} \\
&= -2\frac{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}}\left(\log\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}\frac{\|A\|}{\|x_{k+1}\|} + \frac{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}}{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)}\frac{-2x_{k+1}}{(1-\|x_{k+1}\|)^2}\right)}{\left(\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}\cdot} - 1\right)^2} \\
&\quad + \left(\frac{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} + 1}{\left(\frac{1+\|x_{k+1}\|}{1-\|x_{k+1}\|}\right)^{2\frac{\|Ax_{k+1}\|}{\|x_{k+1}\|}} - 1}\right)\left(\frac{A\|Ax_{k+1}\| - A^TAx_{k+1}}{\|Ax_{k+1}\|^2}\right)
\end{aligned}$$

Now,

$$
\begin{aligned}
\mu_{k+1} &= \frac{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2} \\
\frac{\partial \mu_{k+1}}{\partial x_{k+1}} &= \frac{\partial}{\partial x_{k+1}}\left(\frac{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2}\right) \\
&= \frac{\frac{\partial}{\partial x_{k+1}}(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&- \frac{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)\frac{\partial}{\partial x_{k+1}}(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&= \frac{2\frac{\partial \alpha_{k+1}}{\partial x_{k+1}} z_{k+1}(\|\alpha_{k+1}\|^2\|z_{k+1}\|^2 - \|z_{k+1}\|^2) - 2\frac{\partial \alpha_{k+1}}{\partial x_{k+1}}\|\alpha_{k+1}\|\|z_{k+1}\|^2(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2}.
\end{aligned}
$$

So now we have a closed form for $\dfrac{\partial \mu_{k+1}}{\partial x_{k+1}}$.

Recall that

$$
\nu_{k+1} = \frac{1 - \|\alpha_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2},
$$

then,

$$
\begin{aligned}
\frac{\partial \nu_{k+1}}{\partial x_{k+1}} &= \frac{\partial}{\partial x_{k+1}}\left(\frac{1 - \|\alpha_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2}\right) \\
&= \frac{-2\frac{\partial \alpha_{k+1}}{\partial x_{k+1}}\|\alpha_{k+1}\|\left(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2\right)}{\left(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2\right)^2} \\
&- \frac{(1 - \|\alpha_{k+1}\|^2)\left(-2\frac{\partial \alpha_{k+1}}{\partial x_{k+1}} z_{k+1} + \frac{\partial \alpha_{k+1}}{\partial x_{k+1}} 2\|\alpha_{k+1}\|\|z_{k+1}\|^2\right)}{\left(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2\right)^2}.
\end{aligned}
$$

Now we need the derivatives w.r.t. $z_{k+1}$:

$$
\begin{aligned}
\mu_{k+1} &= \frac{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2} \\
\frac{\partial \mu_{k+1}}{\partial z_{k+1}} &= \frac{\partial}{\partial z_{k+1}}\left(\frac{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2}{1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2}\right) \\
&= \frac{\frac{\partial}{\partial z_{k+1}}(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&- \frac{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)\frac{\partial}{\partial z_{k+1}}(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&= \frac{-2(\alpha_{k+1} - \|z_{k+1}\|)(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&+ 2\frac{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|z_{k+1}\|^2)(\alpha_{k+1} - \|\alpha_{k+1}\|^2\|z_{k+1}\|)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\
&= \frac{2\|z_{k+1}\|(1 - \|\alpha_{k+1}\|\|z_{k+1}\| - 2\langle \alpha_{k+1}, z_{k+1}\rangle - \|\alpha_{k+1}\|^2 + 2\|\alpha_{k+1}\|^2\langle \alpha_{k+1}, z_{k+1}\rangle + \alpha_{k+1}\|z_{k+1}\|)}{(1 - 2\langle \alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2}
\end{aligned}
$$

$$\begin{aligned}\frac{\partial \nu_{k+1}}{\partial z_{k+1}} &= \frac{\partial}{\partial z_{k+1}}\left(\frac{1-\|\alpha_{k+1}\|^2}{1-2\langle\alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2}\right) \\ &= \frac{-\frac{\partial}{\partial z_{k+1}}(1-2\langle\alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)(1-\|\alpha_{k+1}\|^2)}{(1-2\langle\alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2} \\ &= \frac{(2\alpha_{k+1}+2\|\alpha_{k+1}\|^2\|z_{k+1}\|)(\|\alpha_{k+1}\|^2-1)}{(1-2\langle\alpha_{k+1}, z_{k+1}\rangle + \|\alpha_{k+1}\|^2\|z_{k+1}\|^2)^2}.\end{aligned}$$