Appendix: Large-Margin Classification in Hyperbolic Space

A Overview of Other Hyperbolic Space Models

Projecting each point of $\mathbb{L}^n$ onto the hyperplane $x_0 = 0$ using the rays emanating from $(-1, 0, \ldots, 0)$ gives the *Poincaré ball model*

$$\mathbb{B}^n = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : \|x\|^2 < 1\}$$

where the correspondence to the hyperboloid model is given by

$$(x_0, \ldots, x_n) \in \mathbb{L}^n \Leftrightarrow \left(\frac{x_1}{1 + x_0}, \ldots, \frac{x_n}{1 + x_0}\right) \in \mathbb{B}^n.$$

Here, hyperbolic lines are either straight lines that go through the center of the ball or an inner arc of a Euclidean circle that intersects the boundary of the ball at right angles. The Poincaré model is conformal: angles between geodesics in the Poncaré disk are true hyperbolic angles.

If we instead project onto the hyperplane $x_0 = 1$ using rays emanating from the origin, we obtain the *Klein ball model*. The correspondence with the hyperboloid model is given by

$$(x_0, ..., x_n) \in \mathbb{L}^n \Leftrightarrow \left(\frac{x_1}{x_0}, \frac{x_2}{x_0}, ..., \frac{x_n}{x_0}\right) \in \mathbb{K}^n$$

Geodesics in the Klein model are Euclidean straight lines intersecting the unit disk.

Another useful model of hyperbolic space is the *Poincaré half-space model*

$$\mathbb{H}^n = \{\mathbf{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n : x_1 > 0\}$$

which is obtained by taking the inversion of $\mathbb{B}^n$ with respect to a circle that has a radius twice that of $\mathbb{B}^n$ and is centered at a boundary point of $\mathbb{B}^n$. If we center the inversion circle at $(-1, 0, \ldots, 0)$, the resulting correspondence between $\mathbb{B}^n$ and $\mathbb{H}^n$ is given by

$$\begin{aligned}&(x_1, \ldots, x_n) \in \mathbb{B}^n \Leftrightarrow \\ &\frac{1}{1 + 2x_1 + \|x\|^2}\left(1 - \|\mathbf{x}\|^2, 2x_2, \ldots, 2x_n\right) \in \mathbb{H}^n.\end{aligned}$$

In this model, hyperbolic lines are straight lines that are perpendicular to the boundary of $\mathbb{H}^n$ or Euclidean half-circles that are centered on the boundary of $\mathbb{H}^n$.

B Proof of Lemma 4.1

B.1 Proof Based on Hyperbolic Reflection

We thank Reviewer 1 of this paper for providing the idea of the following simplified proof. Our original proof based on Euclidean geometry is included in the subsequent section.

Let $M \in \mathbb{R}^{(n+1)\times(n+1)}$ denote the Minkowski product matrix, with $M_{11} = 1$, $M_{ii} = -1$ for $i > 1$, and $M_{ij} = 0$ for $i \neq j$. Let $\mathbf{w}' = \frac{\mathbf{w}}{\sqrt{-\mathbf{w}*\mathbf{w}}}$, and let $H = I + 2\mathbf{w}'(\mathbf{w}')^T M$. We claim that H corresponds to the hyperbolic reflection through the hyperplane determined by $\mathbf{w} * \mathbf{x} = 0$. To see this, it suffices to show that H preserves both the Minkowski quadratic form, and the hyperplane $\{\mathbf{w} * \mathbf{x} = 0\}$.

Note that

$$\begin{aligned} H^T M H &= (I + 2\mathbf{w}'(\mathbf{w}')^T M)^T M (I + 2\mathbf{w}'(\mathbf{w}')^T M) \\ &= M + 4M^T \mathbf{w}'(\mathbf{w}')^T M + 4M^T \mathbf{w}'(\mathbf{w}')^T M \mathbf{w}'(\mathbf{w}')^T M \\ &= M + 4M^T \mathbf{w}'(\mathbf{w}')^T M - 4M^T \mathbf{w}'(\mathbf{w}')^T M \\ &= M \end{aligned}$$

where in the penultimate equality we use the fact that $(\mathbf{w}')^T M \mathbf{w}' = \mathbf{w}' * \mathbf{w}' = -1$. Thus, H preserves the Minkowski quadratic form.

Suppose that $\mathbf{w} * \mathbf{x} = \mathbf{w}^T M \mathbf{x} = 0$, which implies $(\mathbf{w}')^T M \mathbf{x} = 0$. Then,

$$H\mathbf{x} = (I + 2\mathbf{w}'(\mathbf{w}')^T M)\mathbf{x} = \mathbf{x} + 2\mathbf{w}'((\mathbf{w}')^T M \mathbf{x}) = \mathbf{x}$$

so H preserves the hyperplane.

Now, the hyperbolic distance of $\mathbf{x} \in \mathbb{L}^n$ (i.e., $\mathbf{x} * \mathbf{x} = 1$) to the decision hyperplane defined by $\mathbf{w} * \mathbf{x} = 0$ is just half the hyperbolic distance to the reflection of $\mathbf{x}$ through the hyperplane. To see this, note that the shortest path from $\mathbf{x}$ to its reflection intersects the hyperplane at the closest point on the hyperplane to $\mathbf{x}$. So, the geometric margin can be derived as:

$$\begin{aligned} \frac{1}{2} d^H(\mathbf{x}, H\mathbf{x}) &= \frac{1}{2} \cosh^{-1}(\mathbf{x} * (H\mathbf{x})) \\ &= \frac{1}{2} \cosh^{-1}(\mathbf{x}^T M(\mathbf{x} + 2\mathbf{w}'(\mathbf{w}')^T M\mathbf{x})) \\ &= \frac{1}{2} \cosh^{-1}(\mathbf{x}^T M\mathbf{x} + 2\mathbf{x}^T M\mathbf{w}'(\mathbf{w}')^T M\mathbf{x}) \\ &= \frac{1}{2} \cosh^{-1}(1 + 2(\mathbf{w}' * \mathbf{x})^2) \\ &= \sinh^{-1}(\mathbf{w}' * \mathbf{x}) \\ &= \sinh^{-1}\left(\frac{\mathbf{w} * \mathbf{x}}{\sqrt{-\mathbf{w} * \mathbf{w}}}\right) \end{aligned}$$

where we have used the hyperbolic distance formula $d^H(\mathbf{x}, \mathbf{y}) = \cosh^{-1}(\mathbf{x} * \mathbf{y})$ and trigonometric identity $2\sinh^{-1}(x) = \cosh^{-1}(1 + 2x^2)$. This completes the proof.

B.2 Proof Based on Euclidean Geometry

We aim to derive a closed-form expression for the geometric margin of a given data point $\mathbf{x} \in \mathbb{L}^n$ to a decision hyperplane $\{\mathbf{z} \in \mathbb{L}^n : \mathbf{w} * \mathbf{z} = 0\}$ parameterized by $\mathbf{w} \in \mathbb{R}^{n+1}$.

First, we perform an isometric transformation to simplify calculations. Let A be an orthogonal matrix in $\mathbb{R}^n$. Then, the matrix

$$B = \begin{bmatrix} 1 & 0 \\ 0 & A \end{bmatrix}$$

represents an isometric, orthogonal transformation of the Minkowski space, since it preserves the associated inner product as follows

$$(B\mathbf{u}) * (B\mathbf{v}) = u_0 v_0 - u_{1:n}^T A^T A v_{1:n} = u_0 v_0 - u_{1:n}^T v_{1:n} = \mathbf{u} * \mathbf{v}$$

for any $\mathbf{u}$ and $\mathbf{v}$. If we set the first column of A to $w_{1:n}/\|w_{1:n}\|$ where $\|\cdot\|$ denotes the Euclidean norm, then due to the preservation of geodesics under isometry, the margin of interest becomes equivalent to the margin of a transformed point $\tilde{\mathbf{x}} = Bx$ to the decision hyperplane parameterized by

$$\tilde{\mathbf{w}} = B\mathbf{w} = (w_0, \|w_{1:n}\|, 0, \ldots, 0).$$

The first two coordinates of $\tilde{\mathbf{x}}$ are given in terms of the original coordinates as

$$\tilde{x}_0 = x_0 \text{ and } \tilde{x}_1 = \frac{w_{1:n}^T x_{1:n}}{\|w_{1:n}\|}.$$

Given such a transformation exists for any point in L^n, it is sufficient to derive the margin of an arbitrary point $\tilde{\mathbf{x}} \in L^n$ with respect to a decision hyperplane represented by a weight vector $\tilde{\mathbf{w}}$, where

$$\tilde{\mathbf{w}}_2 = \cdots = \tilde{\mathbf{w}}_n = 0.$$

We use λ to denote the ratio between the first two coordinates of $\tilde{\mathbf{w}}$ as

$$\lambda := \frac{\tilde{w}_0}{\tilde{w}_1}.$$

Importantly, the condition that $\tilde{\mathbf{w}} * \tilde{\mathbf{w}} < 0$ in order for $\tilde{\mathbf{w}}$ to represent a non-trivial decision function is equivalent to the condition that $|\lambda| < 1$, which we will assume in our derivation.

The following lemma characterizes the decision hyperplane defined by such $\tilde{\mathbf{w}}$:

Lemma B.1. *The decision hyperplane $D_{\tilde{\mathbf{w}}} = \{\mathbf{x} : \tilde{\mathbf{w}} * \mathbf{x} = 0, \mathbf{x} \in L^n\}$ corresponding to a weight vector $\tilde{\mathbf{w}} \in \mathbb{R}^{n+1}$ where $\tilde{w}_2 = \cdots = \tilde{w}_n = 0$ is equivalently represented in the Poincaré half-space model as a Euclidean hypersphere centered at the origin with radius $\sqrt{(1-\lambda)/(1+\lambda)}$, where $\lambda = \tilde{w}_0/\tilde{w}_1$.*

Proof. It suffices to show for any $\mathbf{x} \in \mathbb{L}^n$,

$$\mathbf{x} \in D_{\tilde{\mathbf{w}}} \iff \|g(\mathbf{x})\|^2 = \frac{1-\lambda}{1+\lambda},$$

where $g : \mathbb{L}^n \mapsto \mathbb{H}^n$ maps points in the hyperboloid model to the corresponding points in the half-space model and $\|\cdot\|$ denotes the Euclidean norm.

Let $\mathbf{h} \in \mathbb{H}^n$, $\mathbf{b} \in \mathbb{B}^n$, and $\mathbf{x} \in \mathbb{L}^n$ be the points in the half-space model, the ball model, and the hyperboloid model, respectively, that represent the same point in the hyperbolic space. Since

$$h = \frac{1}{1 + 2b_1 + \|\mathbf{b}\|^2}(1 - \|\mathbf{b}\|^2, 2b_2, \ldots, 2b_n),$$

we have

$$\|h\|^2 = \frac{1}{(1 + 2b_1 + \|\mathbf{b}\|^2)^2}\left[(1 - \|\mathbf{b}\|^2)^2 + \sum_{i=2}^{n} 4b_i^2\right].$$

Note that

$$\begin{aligned}(1 - \|\mathbf{b}\|^2)^2 + \sum_{i=2}^{n} 4b_i^2 &= (1 - \|\mathbf{b}\|^2)^2 + 4\|\mathbf{b}\|^2 - 4b_1^2 \\ &= (1 + \|\mathbf{b}\|^2)^2 - 4b_1^2 \\ &= (1 + 2b_1 + \|\mathbf{b}\|^2)(1 - 2b_1 + \|\mathbf{b}\|^2),\end{aligned}$$

which gives us

$$\|h\|^2 = \frac{1 - 2b_1 + \|\mathbf{b}\|^2}{1 + 2b_1 + \|\mathbf{b}\|^2}.$$

Next, recall

$$\mathbf{b} = \frac{1}{x_0 + 1}(x_1, \ldots, x_n),$$

which leads to

$$\|\mathbf{b}\|^2 = \frac{x_1^2 + \cdots + x_n^2}{(x_0+1)^2} = \frac{x_0^2 - 1}{(x_0+1)^2} = \frac{x_0 - 1}{x_0 + 1},$$

where we used the fact that $\mathbf{x} * \mathbf{x} = 1$ since $\mathbf{x} \in \mathbb{L}^n$.

We can now express $\|\mathbf{h}\|^2$ in terms of $\mathbf{x}$ as

$$\|\mathbf{h}\|^2 = \frac{1 - \frac{2x_1}{x_0+1} + \frac{x_0-1}{x_0+1}}{1 + \frac{2x_1}{x_0+1} + \frac{x_0-1}{x_0+1}} = \frac{1 - \frac{x_1}{x_0}}{1 + \frac{x_1}{x_0}}.$$

Because the function $f(z) = (1-z)/(1+z)$ is bijective,

$$\|\mathbf{h}\|^2 = \frac{1-\lambda}{1+\lambda} \iff \frac{x_1}{x_0} = \lambda = \frac{\tilde{w}_0}{\tilde{w}_1}.$$

Finally, note that

$$\begin{aligned}\frac{x_1}{x_0} = \frac{\tilde{w}_0}{\tilde{w}_1} &\iff \tilde{w}_0 x_0 - \tilde{w}_1 x_1 = 0 \\ &\iff \tilde{\mathbf{w}} * \mathbf{x} = 0 \\ &\iff \mathbf{x} \in D_{\tilde{\mathbf{w}}},\end{aligned}$$

where we used the fact that $\tilde{w}_2 = \cdots = \tilde{w}_n = 0$. □

It is a known fact that, in a two-dimensional hyperbolic space, the set of points that are equidistant to a hyperbolic line on the same side of the line forms what is called a *hypercycle*, which takes the shape of a Euclidean circle in the Poincaré half-plane model that goes through the same two *ideal points* as the reference line. Note that the ideal points refer to the two end points of a hyperbolic line in the half-plane model (a circular arc representing a geodesic curve) where the hyperbolic line meets the boundary of the half-plane.

In a high-dimensional setting, an analogous property is that the set of points equidistant to a hyperbolic hyperplane takes the shape of a Euclidean hypersphere that intersects the boundary of the Poincaré half-space at the same ideal points as the hyperplane.

Because our decision hyperplane as characterized in Lemma B.1 has its center at the origin of the half-space model, any hypersphere that intersect the boundary of the half-space at the ideal points of the decision hyperplane must have a center $(c, 0, \ldots, 0)$ for some $c \in \mathbb{R}$. In other words, any hypersphere representing a hypercycle with respect to our given decision boundary is centered on the first coordinate axis (which is perpendicular to the boundary of the half-space).

Let $\tilde{\mathbf{h}} \in \mathbb{H}^n$ be the point in the half-space model that corresponds to the transformed data point $\tilde{\mathbf{x}}$ we described earlier. One way to reason about the margin of $\tilde{\mathbf{x}}$ with respect to the decision hyperplane defined by $\tilde{\mathbf{w}}$ is to find which hypercycle $\mathbf{h}$ belongs to. We can do so using the fact that, in addition to $\tilde{\mathbf{h}}$, an ideal point of the decision hyperplane $(0, r_{\tilde{\mathbf{w}}}, 0, \ldots, 0)$ lies on the hypercycle, where

$$r_{\tilde{\mathbf{w}}} := \sqrt{\frac{1-\lambda}{1+\lambda}}$$

from Lemma B.1. More precisely, we can solve for the center of the hypercycle parameterized by c using the equation

$$(\tilde{\mathbf{h}}_1 - c)^2 + \sum_{i=2}^{n} \tilde{\mathbf{h}}_i^2 = c^2 + r_{\tilde{\mathbf{w}}}^2,$$

which states that $\tilde{\mathbf{h}}$ and $(0, r_{\tilde{\mathbf{w}}}, 0, \ldots, 0)$ are equidistant from the center of the hypercycle $(c, 0, \ldots, 0)$. This gives us

$$c = \frac{\|\tilde{\mathbf{h}}\|^2 - r_{\tilde{\mathbf{w}}}^2}{2\tilde{h}_1}.$$

Now, using the relations

$$\|\tilde{\mathbf{h}}\|^2 = \frac{\tilde{x}_0 - \tilde{x}_1}{\tilde{x}_0 + \tilde{x}_1} \text{ and } \tilde{h}_1 = \frac{1}{\tilde{x}_0 + \tilde{x}_1},$$

which can be derived based on the mapping between the hyperboloid and the half-space models, we obtain that

$$c = \frac{(1 - r_{\tilde{\mathbf{w}}}^2)\tilde{x}_0 - (1 + r_{\tilde{\mathbf{w}}}^2)\tilde{x}_1}{2}.$$

Next, we find a point $(\delta, 0, \ldots, 0) \in \mathbb{H}^n$ that lies on the first coordinate axis and has the same margin with respect to the decision hyperplane as the given data point $\tilde{\mathbf{h}}$.

Since we know that the radius of this hypercycle is given by $\sqrt{c^2 + r_{\tilde{\mathbf{w}}}^2}$ (i.e., the distance from $(c, 0, \ldots, 0)$ to $(0, r_{\tilde{\mathbf{w}}}, 0, \ldots, 0)$), we get

$$\delta = c + \sqrt{c^2 + r_{\tilde{\mathbf{w}}}^2}.$$

Finally, using the hyperbolic distance formula in the half-space model, we obtain that the hyperbolic distance between $(\delta, 0, \ldots, 0)$ and $(r_{\tilde{\mathbf{w}}}, 0, \ldots, 0)$, which is equal to the unsigned geometric margin of our data point, is given by the log-ratio

$$\log \frac{\delta}{r_{\tilde{\mathbf{w}}}} = \log \left(\frac{c}{r_{\tilde{\mathbf{w}}}} + \sqrt{\left(\frac{c}{r_{\tilde{\mathbf{w}}}} \right)^2 + 1} \right) = \operatorname{arsinh} \left(\frac{c}{r_{\tilde{\mathbf{w}}}} \right).$$

Using the expression for c we previously obtained, note that

$$\frac{c}{r_{\tilde{\mathbf{w}}}} = \frac{1}{2} \left[\left(\frac{1}{r_{\tilde{\mathbf{w}}}} - r_{\tilde{\mathbf{w}}} \right) \tilde{x}_0 - \left(\frac{1}{r_{\tilde{\mathbf{w}}}} + r_{\tilde{\mathbf{w}}} \right) \tilde{\mathbf{x}}_1 \right].$$

Since

$$\frac{1}{r_{\tilde{\mathbf{w}}}} - r_{\tilde{\mathbf{w}}} = \sqrt{\frac{1+\lambda}{1-\lambda}} - \sqrt{\frac{1-\lambda}{1+\lambda}} = \frac{2\lambda}{\sqrt{1-\lambda^2}} = \frac{2\tilde{\mathbf{w}}_0}{\sqrt{\tilde{\mathbf{w}}_1^2 - \tilde{w}_0^2}},$$

and similarly,

$$\frac{1}{r_{\tilde{\mathbf{w}}}} + r_{\tilde{\mathbf{w}}} = \frac{2}{\sqrt{1-\lambda^2}} = \frac{2\tilde{w}_1}{\sqrt{\tilde{w}_1^2 - \tilde{w}_0^2}},$$

we can alternatively express the margin as

$$\operatorname{arsinh} \left(\frac{\tilde{w}_0 \tilde{x}_0 - \tilde{w}_1 \tilde{x}_1}{\sqrt{\tilde{w}_1^2 - \tilde{w}_0^2}} \right) = \operatorname{arsinh} \left(\frac{\tilde{\mathbf{w}} * \tilde{\mathbf{x}}}{\sqrt{-\tilde{\mathbf{w}} * \tilde{\mathbf{w}}}} \right),$$

using the fact that $\tilde{w}_2 = \cdots = \tilde{w}_n = 0$.

Finally, since we have shown earlier that our initial transformation of $\mathbf{w}$ and $\mathbf{x}$ to $\tilde{\mathbf{w}}$ and $\tilde{\mathbf{x}}$ preserves the Minkowski inner product, we obtain the geometric margin in terms of the original variables as

$$\operatorname{arsinh} \left(\frac{\mathbf{w} * \mathbf{x}}{\sqrt{-\mathbf{w} * \mathbf{w}}} \right).$$

C Proof of Theorem 4.1

Note that if we make the rescaling $\mathbf{w} \to \kappa \mathbf{w}$, the distance of any point to the decision hyperplane is unchanged. Indeed, using Lemma 4.1, and properties of inner products,

$$\sinh^{-1} \left(\frac{\kappa \mathbf{w} * \mathbf{x}}{\sqrt{-\kappa \mathbf{w} * \kappa \mathbf{w}}} \right) = \sinh^{-1} \left(\frac{\kappa (\mathbf{w} * \mathbf{x})}{\kappa \sqrt{-\mathbf{w} * \mathbf{w}}} \right) = \sinh^{-1} \left(\frac{(\mathbf{w} * \mathbf{x})}{\sqrt{-\mathbf{w} * \mathbf{w}}} \right).$$

Using this freedom, we can assume that $y^0(\mathbf{w} * \mathbf{x}^0) = 1$, where $\mathbf{x}^0$ is the point with the smallest distance to the decision hyperplane specified by $\mathbf{w}$, and $y^0 \in \{-1, 1\}$ is the corresponding decision. This is the analogue of the "canonical representation" of decision hyperplanes familiar from Euclidean SVMs. When a feasible solution $\mathbf{w}$ is thus scaled, we have

$$y^{(j)}(\mathbf{w} * \mathbf{x}^{(j)}) \geq 1$$

for all $j \in [m]$.

Therefore, the optimization problem now simply maximizes

$$\sinh^{-1} \left(\frac{1}{\sqrt{-\mathbf{w} * \mathbf{w}}} \right)$$

which is equivalent to minimizing $-\frac{1}{2} \mathbf{w} * \mathbf{w}$ subject to the above constraint. We add the factor of $\frac{1}{2}$ to simplify gradient calculations. The additional constraint $\mathbf{w} * \mathbf{w} < 0$ ensures that the decision function specified by $\mathbf{w}$ is nontrivial.

D Derivation of Hyperbolic Kernel SVM

D.1 Modified Soft-Margin SVM

When working with hyperbolic kernel SVM, our penalty terms are given by

$$\xi_i = \max(0, 1 - y^i(\mathbf{x}^i * \mathbf{w})). \tag{1}$$

This is conceptually similar to the usual Euclidean penalty, with the Minkowski product in place of the usual Euclidean product. It is heuristically the same in that points on the boundary receive a penalty of 1, points between the margin and the boundary but on the correct side receive a penalty < 1, and points on the wrong side of the boundary receive a penalty > 1. However, unlike in the Euclidean case, the penalty assigned to a point does not scale linearly with its distance from the margin. Indeed, using Theorem 4.1, we see that penalties scale as $\|\mathbf{w}\| \sinh(d^H)$, where d^H denotes hyperbolic distance. This is quite similar to linear scaling for points near the margin, and as we will see, it allows hyperbolic SVM with arbitrary kernels.

We now seek to minimize the objective

$$-\frac{1}{2}\mathbf{w} * \mathbf{w} + C\sum_{i=1}^{m} \xi_i \tag{2}$$

subject to the constraints

$$y^i(\mathbf{w} * \mathbf{x}^i) \geq 1 - \xi_i \tag{3}$$

$$\xi_i \geq 0 \tag{4}$$

$$\mathbf{w} * \mathbf{w} < 0 \tag{5}$$

To parse the first two constraints, we construct a dual formulation. We introduce Lagrange multipliers α_i and μ_i, producing the Lagrangian

$$L(\mathbf{w}, \alpha, \mu) = -\frac{1}{2}\mathbf{w} * \mathbf{w} + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i(\mathbf{w} * \mathbf{x}^i - 1 + \xi_i) - \sum_{i=1}^{m} \mu_i \xi_i. \tag{6}$$

We now compute the stationary values. We see that

$$\frac{\partial L}{\partial w_1} = -w_1 - \sum_{i=1}^{m} \alpha_i (\mathbf{x}^i)_1 y^i. \tag{7}$$

$$\frac{\partial L}{\partial w_j} = w_j + \sum_{i=1}^{m} \alpha_i (\mathbf{x}^i)_j y^i \quad (j > 1) \tag{8}$$

Setting these to zero yields

$$\mathbf{w} = -\sum_{i=1}^{m} \alpha_i y_i \mathbf{x_i}. \tag{9}$$

Likewise, setting the derivatives with respect to ξ_i to zero yields

$$\alpha_i = C - \mu_i. \tag{10}$$

Substituting, and clearing variables using the KKT conditions,we obtain

$$L(\alpha) = -\frac{1}{2}\sum_{i,j=1}^{m} \alpha_i \alpha_j y_i y_j (\mathbf{x^i} * \mathbf{x^j}) + \sum_{i=1}^{m} \alpha_i. \tag{11}$$

Because $\mu_i, \alpha_i \geq 0$ and $\alpha_i = C - \mu_i$, we must have $0 \leq \alpha_i \leq C$. This is the new constraint under which L is minimized. Plugging in the value from equation (9), the negativity constraint $\mathbf{w} * \mathbf{w} < 0$ becomes

$$\sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} (\mathbf{x}^{(i)} * \mathbf{x}^{(j)}) > 0$$

as stated in the main text.

D.2 Polynomial Kernels

The advantage of (11) is that $\mathbf{x}^i * \mathbf{x}^j$ can be replaced with $\phi(\mathbf{x}^i) * \phi(\mathbf{x}^j)$, where ϕ is a feature mapping into the hyperboloid representation of an arbitrarily high-dimensional hyperbolic embedding. Given this, we asked how to construct analogues of the popular polynomial kernel in Euclidean kernel SVM. We cannot simply raise the Minkowski product to a power, as unlike in the Euclidean case, $(\mathbf{x} * \mathbf{y})^d$ is not readily expressed as a product of the monomials of $\mathbf{x}$ and $\mathbf{y}$.

Instead, we endeavored to make the resulting boundary curves analogous. Because hyperbolic space has no natural coordinatization, we rely on the ambient coordinates of the hyperboloid model. A *hyperbolic polynomial curve* in the hyperboloid model with ambient coordinates $x_0, x_1, ..., x_n$ is defined as the intersection of the vanishing set of a homogeneous polynomial $P(x_0, ..., x_n)$ with the forward sheet of the hyperboloid defined by $\mathbf{x} * \mathbf{x} = 1$.

This definition makes intuitive sense on many levels. For example, a *hyperbolic conic* is obtained as the intersection of an elliptic double cone with vertex at the origin with the hyperboloid. In the same way, a standard conic can be defined as the intersection of such a double-cone with a plane.

In the ambient space, the vanishing sets of higher-dimensional homogeneous polynomials are also "cone-like" in the sense that they are closed under scaling of all coordinates: if $P(x_0, ..., x_n) = 0$, then $P(\lambda\, x_0, ..., \lambda\, x_n) = 0$ for any $\lambda \in R$. In particular, these vanishing sets are closed under the gnomonic projection that takes the hyperboloid model to the Klein disk K^m centered at $(1, 0, 0...., 0)$. Therefore, boundary curves produced by polynomial kernels in the hyperboloid model correspond to polynomial planar curves in the Klein model. For example, in the two-dimensional Klein model, hyperbolic conics are intersections of standard ellipses, parabolas, and hyperbolas with the Klein disk.

Thus, the feature mapping corresponding to the polynomial kernel is obtained by first mapping to the Klein disk, then computing the Euclidean polynomial feature mapping in the Klein coordinates, and finally mapping back to the hyperboloid model. Let $\psi_n : \mathbb{R}^{n+1} \to \mathbb{R}^n$ denote the projection that takes the ambient coordinates of the hyperboloid model to the ambient Euclidean coordinates of the Klein model, viewed as the unit disk. One computes that

$$\psi_n(x_0, x_1, ..., x_n) = \left(\frac{x_1}{x_0}, \frac{x_2}{x_0}, ..., \frac{x_n}{x_0}\right). \tag{12}$$

with inverse given by

$$\psi_n^{-1}(y_1, ..., y_n) = \frac{1}{\sqrt{1 - \|y\|^2}}(1, y_1, .y_2, ..., y_n) \tag{13}$$

Let $\phi_k^{n,d} : K^n \to K^{\binom{n+d}{n}}$ be the polynomial feature mapping on the n-dimensional Klein disk, mapping a point $(x_1, ..., x_n)$ to all monomials involving $x_1, ..., x_n$ of degree at most d. Then the feature mapping in hyperbolic space is given by

$$\phi_h = (\psi_{\binom{n+d}{n}})^{-1} \circ \phi_k^{n,d} \circ \psi_n \tag{14}$$

The corresponding kernel is:

$$K_h(\mathbf{x}, \mathbf{y}) = \frac{K_k(\mathbf{x}, \mathbf{y}) - 1}{\sqrt{(1 - K_k(\mathbf{x}, \mathbf{x}))(1 - K_k(\mathbf{y}, \mathbf{y}))}} \tag{15}$$

where $K_k(x, y) = (1 + \psi_n(\mathbf{x}) \cdot \psi_n(\mathbf{y}))^d$ is the polynomial kernel in Klein coordinates.

D.3 Proof of Theorem 5.1

The backward direction of the theorem requires the following Lemma:

Lemma D.1. *Let $\boldsymbol{x}$ and $\boldsymbol{y}$ be two points on the hyperbola H defined by $\{\boldsymbol{z} : \boldsymbol{z} * \boldsymbol{z} = 1\}$ in $\mathbb{R}^n$. If $\boldsymbol{x}$ and $\boldsymbol{y}$ lie on the same sheet of $\boldsymbol{H}$, then $\boldsymbol{x} * \boldsymbol{y} \geq 1$. Otherwise, $\boldsymbol{x} * \boldsymbol{y} \leq -1$.*

Proof. Suppose first that $\mathbf{x}$ and $\mathbf{y}$ lie on the same sheet of H. Then $\mathbf{x}_0$ and $\mathbf{y}_0$ have the same sign. By the Cauchy-Schwarz inequality, we have

$$\mathbf{x} * \mathbf{y} = \sqrt{(1 + \|\mathbf{x}_{1:n}\|^2)(1 + \|\mathbf{y}_{1:n}\|^2)} - \mathbf{x}_{1:n} \cdot \mathbf{y}_{1:n} \geq \sqrt{(1 + \|\mathbf{x}_{1:n}\|^2)(1 + \|\mathbf{y}_{1:n}\|^2)} - \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\|$$
$$\geq 1 + \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\| - \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\| = 1.$$

For the last inequality note that, by the inequality of arithmetic and geometric means,

$$(1 + \|\mathbf{x}_{1:n}\|^2)(1 + \|\mathbf{y}_{1:n}\|^2) = 1 + \|\mathbf{x}_{1:n}\|^2 \, \|\mathbf{y}_{1:n}\|^2 + \|\mathbf{x}_{1:n}\|^2 + \|\mathbf{y}_{1:n}\|^2$$
$$\geq 1 + \|\mathbf{x}_{1:n}\|^2 \, \|\mathbf{y}_{1:n}\|^2 + 2 \, \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\| = (1 + \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\|)^2.$$

If $\mathbf{x}$ and $\mathbf{y}$ lie on different sheets of the hyperbola, then $\mathbf{x}_0$ and $\mathbf{y}_0$ have different signs, so the first term of the Minkowski product is negative. By Cauchy-Schwarz and the above argument, we have:

$$\mathbf{x} * \mathbf{y} = -\sqrt{(1 + \|\mathbf{x}_{1:n}\|^2)(1 + \|\mathbf{y}_{1:n}\|^2)} - \mathbf{x}_{1:n} \cdot \mathbf{y}_{1:n}$$
$$\leq -\big(\sqrt{(1 + \|\mathbf{x}_{1:n}\|^2)(1 + \|\mathbf{y}_{1:n}\|^2)} + \|\mathbf{x}_{1:n}\| \, \|\mathbf{y}_{1:n}\| \, \big) \leq -1$$

as desired. □

We are now ready to prove Theorem 5.1.

Proof of Theorem 5.1. **Forward direction**: Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \in L^n$ and $\mathbf{M} \in \mathbb{R}^{m \times m}$ such that $M_{ij} = -\mathbf{x}_i * \mathbf{x}_j$. Define

$$\mathbf{z}_i = \begin{bmatrix} x_{1i} & \cdots & x_{mi} \end{bmatrix}^T$$

for $i = 0, \ldots, n$ and

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1 & \cdots & \mathbf{z}_n \end{bmatrix}.$$

Then we can represent M as

$$\mathbf{M} = \mathbf{Z}\mathbf{Z}^T - \mathbf{z}_0\mathbf{z}_0^T.$$

To show $\mathbf{M}$ has exactly one negative eigenvalue, it is sufficient to prove the following two statements: (i) $\mathbf{M}$ has at most one negative eigenvalue, and (ii) $\mathbf{M}$ is not positive semidefinite (PSD). Note that (ii) rules out the possibility of $\mathbf{M}$ having zero negative eigenvalues.

For (i), we first note that $\mathbf{Z}\mathbf{Z}^T$ can be eigendecomposed to $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ with an orthogonal matrix $\mathbf{Q}$ and a diagonal matrix $\mathbf{\Lambda}$ of non-negative eigenvalues $0 \leq \lambda_1 \leq \cdots \leq \lambda_m$, where the non-negativity follows from the fact $\mathbf{Z}\mathbf{Z}^T$ is PSD. Now, consider the matrix

$$\mathbf{M}' := \mathbf{Q}^T\mathbf{M}\mathbf{Q} = \mathbf{\Lambda} - (\mathbf{Q}^T\mathbf{z}_0)(\mathbf{Q}^T\mathbf{z}_0)^T.$$

Because $\mathbf{M}'$ is obtained by orthogonal transformations of $\mathbf{M}$, the two matrices $\mathbf{M}$ and $\mathbf{M}'$ share the same set of eigenvalues. Note $\mathbf{M}'$ is obtained by updating a diagonal matrix with a rank-one matrix. The impact of such a rank-one update on eigenvalues of a matrix has been studied by Golub (1973). In particular, if we denote the eigenvalues of $\mathbf{M}'$ with $\lambda'_1 \leq \cdots \leq \lambda'_m$, it is known that

$$\lambda'_1 \leq \lambda_1 \leq \lambda'_2 \leq \cdots \leq \lambda'_m \leq \lambda_m.$$

Since $\lambda_1 \geq 0$, we conclude that $\mathbf{M}$ has at most one negative eigenvalue.

Next we prove (ii) by contradiction. Assume $\mathbf{M}$ is PSD, which implies $\mathbf{u}^T\mathbf{M}\mathbf{u} \geq 0$ for all $\mathbf{u} \in \mathbb{R}^m$. For any $\mathbf{u} \in \mathbf{null}(\mathbf{Z}^T)$, we observe that

$$0 \leq \mathbf{u}^T\mathbf{M}\mathbf{u} = \mathbf{u}^T\mathbf{Z}\mathbf{Z}^T\mathbf{u} - \mathbf{u}^T\mathbf{z}_0\mathbf{z}_0^T\mathbf{u} = -(\mathbf{u}^T\mathbf{z}_0)^2 \leq 0,$$

which gives $\mathbf{u}^T\mathbf{z}_0 = 0$. Since this implies $\mathbf{z}_0 \perp \mathbf{null}(\mathbf{Z}^T)$, we conclude $\mathbf{z}_0 \in \mathbf{range}(\mathbf{Z})$, and thus there exists $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{Z}\mathbf{w} = \mathbf{z}_0$.

Since $\mathbf{w}' := \mathbf{w} + \mathbf{v}$ for any $\mathbf{v} \in \mathbf{null}(\mathbf{Z})$ also satisfies $\mathbf{Z}\mathbf{w}' = \mathbf{z}_0$, we can assume $\mathbf{w} \in \mathbf{range}(\mathbf{Z}^T)$ without loss of generality. That is, there exists $\mathbf{s} \in \mathbb{R}^m$ such that $\mathbf{w}^T = \mathbf{s}^T\mathbf{Z}$.

Recall that $\mathbf{Z}$ and $\mathbf{z}_0$ represent the coordinates of data points in the hyperboloid model. Given any data point $\mathbf{x}_i = [x_{i0}, \ldots, x_{in}]^T$ and letting $\mathbf{x}_{i1:n}$ denote the last n coordinates, the above linear dependence of x_{i0} on $\mathbf{x}_{i1:n}$ (from $\mathbf{Z}\mathbf{w} = \mathbf{z}_0$) implies

$$\|\mathbf{x}_{i1:n}\|^2 + 1 = x_{i0}^2 = (\mathbf{w}^T \mathbf{x}_{i1:n})^2 \leq \|\mathbf{w}\|^2 \cdot \|\mathbf{x}_{i1:n}\|^2,$$

where the first equation follows from $\mathbf{x}_i \in L^n$, and the last inequality is given by Cauchy-Schwarz. Thus, $\|\mathbf{w}\| > 1$.

Finally, we invoke the definition of PSD for $\mathbf{M}$ with the previously defined vector $\mathbf{s}$ (which satisfies $\mathbf{w}^T = \mathbf{s}^T\mathbf{Z}$) to obtain

$$0 \leq \mathbf{s}^T\mathbf{M}\mathbf{s} = \mathbf{s}^T(\mathbf{Z}\mathbf{Z}^T - \mathbf{z}_0\mathbf{z}_0^T)\mathbf{s} = \mathbf{s}^T\mathbf{Z}(\mathbf{I} - \mathbf{w}\mathbf{w}^T)\mathbf{Z}^T\mathbf{s} = \|\mathbf{w}\|^2 - \|\mathbf{w}\|^4,$$

which is a contradiction given $\|\mathbf{w}\| > 1$. This completes the proof for (ii).

Backward direction: Because M is symmetric, it has an eigendecomposition PDP^T, where the columns of P form an orthonormal eigenbasis of $\mathbb{R}^n$ with respect to M, and D is diagonal with eigenvalue entries $\lambda_0, ..., \lambda_n$. Assume without loss of generality that λ_0 is the negative eigenvalue.

Re-write M as $M = QQ^T$, where $Q = PD^{\frac{1}{2}}$. The diagonal entries of $D^{\frac{1}{2}}$ are $\sqrt{\lambda_0}, ..., \sqrt{\lambda_n}$, and because λ_0 is negative, $\sqrt{\lambda_0} = i\sqrt{-\lambda_0}$. Thus, the first column of Q inherits a factor of i. Let $\tilde{\mathbf{x}_1}, ..., \tilde{\mathbf{x}_n}$ denote the rows of Q, so that each $\tilde{\mathbf{x}_i}$ has an imaginary first coordinate. Obtain $\phi(\mathbf{x}_i)$ from $\tilde{\mathbf{x}_i}$ by removing the factor of i from the first coordinate. Note that

$$M_{ij} = \tilde{\mathbf{x}_i} \cdot \tilde{\mathbf{x}_j} = -(\phi(\mathbf{x}_i) * \phi(\mathbf{x}_j))$$

as desired.

All that remains is to ensure that $\phi(\mathbf{x}_i)$ and $\phi(\mathbf{x}_j)$ lie on the same sheet of the hyperboloid defined by $\mathbf{x} * \mathbf{x} = 1$. Because the diagonals of M are all -1, we have $\phi(\mathbf{x}_i) * \phi(\mathbf{x}_i) = 1$ for all i, so the feature vectors lie on the hyperboloid. Because the remaining entries of M are at most -1, $\phi(\mathbf{x}_i) * \phi(\mathbf{x}_j) \geq 1$ for all i, j. By Lemma D.1, our constructed feature vectors $\phi(\mathbf{x}_i)$ all lie on the same sheet of a hyperboloid, so they may be viewed as embedded in the hyperboloid model of hyperbolic space.

□

D.4 Implementation Details of Quadratic Hyperbolic SVM

The precise form of quadratic hyperbolic SVM used in our experiments is obtained by applying the bootstrap technique of Lemma 5.1 to the following Euclidean kernel

$$k_E(\mathbf{x}, \mathbf{y}) = \frac{(\mathbf{x}^T\mathbf{y})^2 + \mathbf{x}^T\mathbf{y}}{2},$$

which indeed satisfies the condition that $k_E(\mathbf{x}, \mathbf{x}) < 1$ for all $\|\mathbf{x}\| < 1$. Intuitively, this corresponds to the class of decision functions that can be expressed as the intersection of the hyperboloid model with a polynomial consisting of monomials only up to degree 2. Due to the extra dimension in the ambient space of hyperboloid model, the corresponding Euclidean kernel with the same degrees of freedom in the Euclidean setting is $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y}+1)^2$, which additionally leverages an intercept term. As baseline in our comparisons, we solve the Euclidean kernel SVM with the $(\mathbf{x}^T\mathbf{y}+1)^2$ kernel using the LIBSVM package (`https://www.csie.ntu.edu.tw/~cjlin/libsvm/`). For our method, although one may directly solve the kernel formulation of hyperbolic SVM (Section 5.1), we observed that existing methods for solving SVMs with indefinite kernels empirically lack robustness. We thus opted instead to optimize the primal problem with the equivalent feature map using our hyperbolic linear SVM implementation.

E Euclidean SVMs in Other Hyperbolic Space Models

Although Euclidean SVMs in our main experiments were based on data points in the Poincaré ball model, in principle, Euclidean SVM can be applied to data points represented in any hyperbolic space model. Figure E.1 shows additional comparisons of hyperbolic SVM with Euclidean SVM where the input coordinates are given in the hyperboloid and Klein models. In both cases, hyperbolic SVM significantly outperforms Euclidean SVM with one-sided paired-sample t-test p-values of 7.20×10^{-9} for the hyperboloid model and 1.58×10^{-25} for

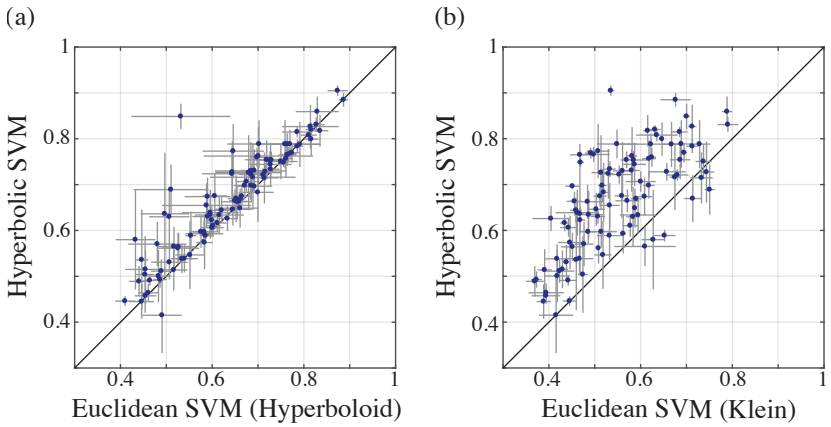

Figure E.1: **Comparison of hyperbolic and Euclidean SVMs based on the hyperboloid and Klein models.** Panels (a) and (b) correspond to the experiments depicted in Figure 2a, but with input data coordinates for Euclidean SVM provided in the hyperboloid and Klein models, respectively. In addition, as in Figure 4a, we used a mixture of elliptical Gaussian distributions instead of isotrophic ones to evaluate the methods based on a more challenging setting. Hyperbolic SVM significantly outperforms both versions of Euclidean SVMs.

the Klein model. Note that in the Klein model, hyperbolic geodesics coincide with the geodesics of the ambient Euclidean space. Thus, Euclidean SVM based on the Klein model considers the same class of decision boundaries as hyperbolic SVM, yet achieves poor performance due to the distortions introduced in margin calculations.