SPACETIME REPRESENTATION LEARNING

Marc T. Law & James Lucas
NVIDIA

ABSTRACT

Much of the data we encounter in the real world can be represented as directed graphs. In this work, we introduce a general family of representations for directed graphs through connected time-oriented Lorentz manifolds, called “spacetimes” in general relativity. Spacetimes intrinsically contain a causal structure that indicates whether or not there exists a causal or even chronological order between points of the manifold, called events. This chronological order allows us to naturally represent directed edges via imposing the correct ordering when the nodes are embedded as events in the spacetime. Previous work in machine learning only considers embeddings lying on the simplest Lorentz manifold or does not exploit the connection between Lorentzian pre-length spaces and directed graphs. We introduce a well-defined approach to map data onto a general family of spacetimes. We empirically evaluate our framework in the tasks of hierarchy extraction of undirected graphs, directed link prediction and representation of directed graphs.

1 INTRODUCTION

Most of the machine learning literature has focused on learning representations that lie on a Riemannian manifold such as the Euclidean space, the $d$-sphere (e.g., $l_2$-normalized representations) (Wang et al., 2017; Tapaswi et al., 2019), hyperbolic geometry to represent graphs without cycles (Nickel & Kiela, 2017), or a statistical manifold in information geometry (Amari, 1998). Concepts of Euclidean geometry, such as distances, are naturally generalized to Riemannian geometry which remains easy to interpret. In contrast, recent approaches have considered learning representations that lie on a pseudo-Riemannian manifold to extract hierarchies in graphs with cycles (Law & Stam, 2020; Law, 2021) or represent directed graphs (Clough & Evans, 2017; Sim et al., 2021).

Pseudo-Riemannian manifolds are generalizations of Riemannian manifolds where the constraint of positive definiteness of the nondegenerate metric tensor is relaxed. The machine learning literature on pseudo-Riemannian manifolds can be divided into two categories. The first category focuses on how to optimize a given function whose domain is a pseudo-Riemannian manifold and does not take into account whether the manifold is time-oriented or not (Law & Stam, 2020; Law, 2021). The second category exploits the interpretation of a specific family of pseudo-Riemannian manifolds called “spacetimes” in general relativity (Clough & Evans, 2017; Sim et al., 2021). More specifically, spacetimes are connected time-oriented Lorentz manifolds. They intrinsically contain a causal structure that indicates whether or not there exists a causal order between points of the manifold, called events. This causal structure has been utilized to represent directed graphs where each node is an event and the existence of an arc (i.e., directed edge) between two nodes depends on the causal character of the curves joining them (Bombelli et al., 1987). In particular, Clough & Evans (2017) consider learning representations via the Minkowski spacetime which is the simplest such manifold. On the other hand, Sim et al. (2021) use three types of spacetimes and propose an ad hoc method based on the sign of some time coordinate difference function to determine the orientation of edges. The sign of such a function is not always meaningful as it for instance alternates periodically when the manifold is non-chronological and does not generalize to all spacetimes. Moreover, the distance function that they optimize is constant when two points cannot be joined by a geodesic.

Contributions. We propose a framework inspired by Lorentzian causality theory (Kronheimer & Penrose, 1967; Minguzzi, 2019), and in particular Lorentzian pre-length spaces (Kunzinger & Sämann, 2018), to learn directed graph representations lying on a large family of spacetimes. To this end, we present tools to account for time-orientation and exploit distances specific to Lorentz geometry. In particular, we propose to restrict the existence of edges to pairs of nodes whose
We introduce some differential geometry background about spacetimes. Spacetimes have been widely studied, and we refer the reader to Hawking & Ellis (1973), Uhlenbeck (1975), O’Neill (1993), Beem et al. (1996); Wolf (2011); Gourgoulhon (2016), Chapter 5-8 & 14 of O’Neill (1983).

Pseudo-Riemannian Manifold. A $d$-dimensional pseudo-Riemannian manifold $(M, g)$ is a smooth manifold such that every point $x \in M$ has a $d$-dimensional tangent space $T_xM$ whose metric tensor $g_x : T_xM \times T_xM \to \mathbb{R}$ is a nondegenerate symmetric bilinear form (called a scalar product). Nondegeneracy means that $\forall \textbf{u} \in T_xM$, $g_x(\textbf{u}, \textbf{u}) = 0 \implies \textbf{u} = 0$. When the context is clear and to simplify the notation, we write $\langle \cdot, \cdot \rangle := g_x(\cdot, \cdot)$ to define the metric tensor at $x$. We also write $M$ instead of $(M, g)$. We write points $x \in M$ of the manifold in bold serif font, and tangent vectors $\textbf{u} \in T_xM$ in bold sans-serif font when we want to distinguish them from points.

Lorentz manifold. Every tangent space $T_xM$ of a $d$-dimensional pseudo-Riemannian manifold $M$ admits an orthonormal basis $\{\textbf{e}_1, \ldots, \textbf{e}_d\}$ that satisfies $\forall i, \langle \textbf{e}_i, \textbf{e}_j \rangle = \pm 1$ and $\forall i \neq j, \langle \textbf{e}_i, \textbf{e}_j \rangle = 0$. The index $\nu \leq d$ of $M$ is the number of vectors $\textbf{e}_i$ that satisfy $\langle \textbf{e}_i, \textbf{e}_i \rangle = -1$. If $\nu = 0$, $M$ is Riemannian and its metric tensor is positive definite (i.e., $\forall x \in M, \forall \textbf{u} \in T_xM, \langle \textbf{u}, \textbf{u} \rangle \geq 0$ and $\langle \textbf{u}, \textbf{u} \rangle = 0 \iff \textbf{u} = 0$). If $\nu = 1$, $M$ is a Lorentz manifold and $T_xM$ is a Lorentz vector space.

Future timecone. A nonzero tangent vector $\textbf{u}$ is called timelike (or chronological), null, spacelike or non-spacelike (or causal) if $\langle \textbf{u}, \textbf{u} \rangle$ is negative, zero, positive or nonpositive, respectively. The type into which $\textbf{u}$ falls is called its causal character. If $\textbf{u} = 0$, then $\textbf{u}$ is spacelike. Every Lorentz tangent space contains two timecones. Some timelike tangent vector $\textbf{t} \in T_xM$ can arbitrarily be used to define the future timecone as the following set: $C^+_f(\textbf{t}) := \{\textbf{v} \in T_xM : \langle \textbf{v}, \textbf{v} \rangle < 0, \langle \textbf{t}, \textbf{v} \rangle < 0\}$ whereas $-\textbf{t}$ defines the past timecone $C^-_f(\textbf{t}) := C^+_f(-\textbf{t})$. Two timelike tangent vectors $\textbf{u}$ and $\textbf{v}$ are in the same timecone iff $\langle \textbf{u}, \textbf{v} \rangle < 0$. They belong to different timecones if $\textbf{v} = -\textbf{u}$.

Time-orientability and time-orientation. A continuous vector field $X$ is a function that assigns to each point $x \in M$ a tangent vector of $M$ at $x$ denoted by $X(x) \in T_xM$. $X$ and $-X$ are timelike if $\forall x \in M, \langle X(x), X(x) \rangle < 0$. A Lorentz manifold is time-orientable iff there exists a timelike vector field. If $M$ is assigned such a timelike vector field $X$, it is time-oriented by $X$. In this case, non-spacelike tangent vectors $\textbf{u}$ at each point $x$ can be divided into two separate classes: future-directed if $\langle X(x), \textbf{u} \rangle < 0$, and past-directed if $\langle X(x), \textbf{u} \rangle > 0$.

A curve $\gamma_{x\rightarrow y} : I \to M$ where $I \subseteq \mathbb{R}$ is defined such that its initial point is $\gamma_{x\rightarrow y}(0) = x$ and its initial velocity is $\gamma'_{x\rightarrow y}(0) = \textbf{u} \in T_xM$. We denote it by $\gamma$ when its initial conditions are clear from
the context. If its acceleration is zero, then it is called a geodesic. The curve γ is called timelike, null or spacelike if its velocity on the whole domain I is timelike, null or spacelike, respectively. It is called future-directed (or future-pointing) if its velocity is future-directed (and causal) on I.

**Completeness.** The manifolds that we consider in this paper are geodesically complete (i.e., $I = \mathbb{R}$), and the exponential map $\exp_x : T_x M \to M$ of $M$ at $x$ is defined as $\exp_x(u) := \gamma_{x \to u}(1)$ where $\gamma_{x \to u}$ is a geodesic. The maximal normal neighborhood of $x$ is the maximal subset $U_x \subseteq M$ where the logarithmic map $\log_x : T_x M \to T_x \mathbb{R}^d$, $\log_x := \exp_x^{-1}$, is a diffeomorphism, satisfies $U_x := \{ y \in M : \exp_x(\exp_x^{-1}(y)) = y \}$. To simplify the notation, we also write $\tilde{x}^y := \log_x(y)$ where $y \in U_x$. $U_x$ is convex if $\forall y \in U_x$, there exists a unique geodesic totally contained in $U_x$ from $x$ to $y$.

We now present some pseudo-Riemannian manifolds that will be relevant. Their differential geometry tools (e.g., exponential/logarithmic map and parallel transport) are provided in Appendix B.

- **Pseudo-Euclidean space** $\mathbb{R}^d_\nu$. The flat $d$-dimensional pseudo-Riemannian manifold of index $\nu$ is denoted by $\mathbb{R}^d_\nu$. In particular, $\mathbb{R}^d_0$ is the Euclidean space and $\mathbb{R}^d_1$ is the Minkowski space. Since it is a vector space, we can identify its tangent space to the space itself by means of the natural isomorphism $\mathbb{R}^d_\nu \approx T_x \mathbb{R}^d_\nu$. $\mathbb{R}^d_\nu$ is equipped with the following scalar product:

$$\forall x = (x_1, \ldots, x_d)^T, \quad y = (y_1, \ldots, y_d)^T, \quad (x, y)_\nu := - \sum_{i=1}^d x_i y_i + \sum_{j=1}^d x_j y_j. \quad (1)$$

The maximal normal neighborhood for all $x \in \mathbb{R}^d_\nu$ is $U_x = \mathbb{R}^d_\nu$. In special relativity, the first $\nu$ elements of $x \in \mathbb{R}^d_\nu$ are time coordinates and the other ones are called space coordinates.

- **The pseudo-sphere** $S^d_\nu(r)$ of radius $r > 0$ is called the de Sitter space when $\nu = 1$ and defined as:

$$S^d_\nu(r) := \{ x \in \mathbb{R}^d_{\nu+1} : (x, x)_\nu = r^2 \}. \quad (2)$$

It is not time-orientable if $d - \nu$ is odd, and $U_x = \{ y \in S^d_\nu(r) : (x, y)_\nu > -r^2 \}$.

- **The pseudo-hyperbolic space** $H^d_\nu(r)$ is called the anti-de Sitter space when $\nu = 1$ and defined as:

$$H^d_\nu(r) := \{ x \in \mathbb{R}^d_{\nu+1} : (x, x)_{\nu+1} = -r^2 \}. \quad (3)$$

The anti-de Sitter space is time-orientable for all $d$, and $U_x = \{ y \in H^d_\nu(r) : (x, y)_{\nu+1} < r^2 \}$.

- **The cylindrical Minkowski space** $L^d_\nu(C) := \mathbb{R}^d_{\nu}/\sim$ is a quotient set defined such that $x \in \mathbb{R}^d_\nu$ and $y \in \mathbb{R}^d_\nu$ are equivalent (i.e., $x \sim y$) if $\forall i > 0, y_i = x_i$ and $\exists k \in \mathbb{Z}, y_0 = x_0 + kC$ where $C > 0$ is a circumference hyperparameter. See page 148 of [O'Neill](1983) for other types of Lorentz cylinders. We have $U_x = \{ y = (y_0, \ldots, y_{d-1})^T \in \mathbb{R}^d_1 : y_0 \in (x_0 - C/2, x_0 + C/2) \}$.

3 **Spacetime Graph Representation**

A **spacetime** is a connected time-oriented Lorentz manifold $M$ whose points are called events. Informally, *time-oriented* is often weakened to *time-orientable*, and [Hawking & Ellis](1973) ignore the time-orientability criterion to define spacetimes although it is required to study their causal structure. Our main contribution is to exploit this causal structure via Lorentzian pre-length spaces to represent graphs. In the following, we always assume that $M$ is a spacetime unless stated otherwise. A finite directed graph can be given the structure of a Lorentzian pre-length space ([Kunzinger & Sämann](2018)). To define our graphs, we then choose a special type of Lorentzian pre-length space that is easy to optimize. We now provide and follow the definitions of [Kunzinger & Sämann](2018).

**Causal space.** Let $\mathcal{X}$ be a set with a reflexive and transitive relation $\leq$, and $\ll$ a transitive relation contained in $\leq$ (i.e., $x \ll y \implies x \leq y$). Then $(\mathcal{X}, \ll, \leq)$ is called a causal space. This definition is more general than the one given in the seminal work of [Kronheimer & Penrose](1967).

Following general relativity, the event $x \in M$ causally (resp. chronologically) precedes the event $y \in M$, and we write $x \ll y$ (resp. $x \ll y$) if there exists a future-directed causal (resp. timelike) curve from $x$ to $y$. This condition might be difficult to verify in general. However, if $y$ is in a convex normal neighborhood of $x$ denoted by $\mathcal{V}_x \subseteq U_x$ (and we always assume in the following that $x \in \mathcal{V}_x$), we have $x \ll y$ (resp. $x \ll y$) iff there exists a nonconstant future-directed causal (resp. timelike) geodesic from $x$ to $y$ (see Proposition 4.5.1 of [Hawking & Ellis](1973)). We note
x ≤ y ⇐⇒ x = y or x < y. Any open subset W ⊆ M of a spacetime is a spacetime in its own right, and the intrinsic causal relations of W imply the corresponding ones in M. In the case where W ⊆ M is a convex open set, the intrinsic causality of W is as simple as that of Minkowski space (see page 403 of O'Neill (1983)). Moreover, both (M, ≪, ≤) and (W, ≪, ≤) are causal spaces.

Lorentzian pre-length space. Let (X, ≪, ≤) be a causal space and d a metric on X. Let τ : X × X → [0, ∞] be a lower semicontinuous map (w.r.t. the metric topology induced by d) that satisfies the reverse triangle inequality: ∀x, y, z ∈ X with x ≤ y ≤ z, τ(x, z) ≥ τ(x, y) + τ(y, z). Suppose that τ(x, y) = 0 if x ≪ y and τ(x, y) > 0 iff x ≪ y. Then (X, ≪, ≤) is a Lorentzian pre-length space and τ is called the time separation function. See details about τ in Section 3.3.

3.1 Graph construction via Lorentzian pre-length spaces

We now explain our methodological contribution, which is the construction of directed graphs from the properties mentioned above. Let us define a directed graph G = (V, E) where V = {vi}i∈I is the node set and E is the set of arcs. Our goal is to represent each node vi by a point xi ∈ M so that there exists an arc from vi to vj (i.e., (vi, vj) ∈ E) only if xi ≪ xj. We propose to consider subgraphs Gi = (V, Ei) defined such that E = ∪i∈I Ei. Each subgraph Gi is given by the structure of the Lorentzian pre-length space (Vxi, d, ≪, τ) where Vxi, xi ∈ M is an open subset of the maximal normal neighborhood Uxi. More precisely, we consider the chronological future I+(xi, Vxi) := {y ∈ Vxi : xi ≪ y} of the point xi relative to some given set Vxi ⊆ Uxi (see page 402 of O'Neill (1983)), and we draw an arc from vi to vj iff xj ∈ I+(xi, Vxi). Assuming Vxi is a causal space and xj ∈ Vxi, we have xj ∈ I+(xi, Vxi) only if xi ≺ xj is defined and future-directed timelike (see Proposition 4.12 of Beem et al. (1996)).

Our framework produces a subgraph G = ∪i∈I Gi of the graph G described by the chronological relations of M, and the causal relations between events depend on the choice of M. It is worth noting that for any spacetime, chronological order is transitive (i.e., xi ≺ xj and xj ≺ xk ⇒ xi ≺ xk (i.e., xk ∈ I+(xi, M)), see Chapter 3.2 of Beem et al. (1996)). vi is then connected by an arc to all the successors of vi in G. We avoid this degenerate case by drawing an arc from vi to vj in G iff xk ∈ I+(xi, Vxi). The choice of M and Vxi ⊆ M for all i is then crucial to construct G ⊆ G.

Existence of directed cycles. Our framework can represent graphs with directed cycles only if M is non-chronological (i.e., there exists at least one closed timelike curve: ∃x ∈ M, x ≺ x, see Fig. 1], which is the case if M is the anti-de Sitter space Sd+1(r) or the Cylindrical Minkowski space C×(C). Spacetimes that do not contain closed timelike curves are called chronological (i.e., ∃x ∈ M, x ≺ x) and can represent only Directed Acyclic Graphs (DAGs) in our framework. Some chronological spacetimes such as Sd+1(r) (when d ≥ 3 is odd) or Sd+1 are called globally hyperbolic and satisfy: x ≺ y =⇒ x and y can be joined by a (longest) causal geodesic that is not necessarily unique (see page 66 of Beem et al. (1996)). Nonetheless, if M is Sd+1(r) or Sd+1, we also have y ∈ I+(x, M) iff y ∈ Ux and xy is future-directed timelike (see page 411 of O'Neill (1983) for details on conditions). Although DAGs do not contain directed cycles, they can contain undirected cycles.

For simplicity, we consider that Vxi ⊆ Uxi is the convex normal neighborhood of x that contains points y such that the arc length dγ of the geodesic γ from x = γ(0) to y = γ(1) is smaller than some arbitrary threshold ε ∈ (0, ∞] and can be formulated as dγ(x, y) := √(xy, xy). We have:

\[ I^+(x, Vx) = \{ y \in Vx : -\varepsilon^2 < \langle xy, xy \rangle < 0, xy \in C^y_x (t) \}, \]

where C^y_x (t) is the future timecone parametrized by some arbitrary timelike tangent vector t ∈ TxM. The motivation is to choose or learn ε small enough that vi is not connected to undesired successors of vj. However, in most of our experiments we simply consider that Vxi = Uxi (see Section C.4).
Therefore for many spacetimes, there exist pairs of points that cannot be joined by a geodesic even if the spacetime is geodesically complete. Working with convex normal neighborhoods $V_x$, allows us to constrain chronological order between points only via timelike geodesics as explained in Section 3.1.

- **Minkowski spacetime** $\mathbb{R}^d$. We recall that $\forall \nu, \mathbb{R}_\nu^d \approx T_x \mathbb{R}^d$. Its geodesic $\gamma_{\mathbb{R}^d}(t) : x + ty$. The exponential map at $x$ is $\exp_x(y) := x + y$ and its inverse is $\chi_y := y - x$. $\mathbb{R}^d$ is time-oriented by the vector field $\partial / \partial x_0$ (e.g., $\forall x, \tau(x, y) := y_0 - x_0$, and we should in theory define $\tau(x, y) := 0$ if $x \not\in y$ to properly follow the definition of Lorentzian pre-length spaces but we ignore this criterion during training for optimization purpose). Let us define $t := (1, 0, \ldots, 0)^T$, $\alpha = -(y_0 - x_0) + \sqrt{\sum_{i=1}^{d-1} (y_i - x_i)^2}$ and $\beta := (\chi_y, \chi_y) := -(y_0 - x_0)^2 + \sum_{i=1}^{d-1} (y_i - x_i)^2$.

According to equation 4 we have $\exists \gamma \in \gamma_{\mathbb{R}^d}(t)$ such that $\gamma(\tau) = x$. If $x \subset x$ (i.e., $x \subset y$) then $\gamma(\tau) = x$.

- **de Sitter spacetime** $\mathbb{H}^d(r)$. The original de Sitter spacetime $\mathbb{H}^d(r)$ is not time-orientable (see Corollary 11.2.6 of [Wolf 2011]). However, when $d \geq 3, \nu > 0$ and $d - \nu$ is even, $\mathbb{H}^d(r)$ is orientable and timelike, the projective space $\mathbb{H}^d_p(r) \pm 1$ used in [Law 2021] is orientable but not time-orientable (see page 247 of [O'Neill 1983]). We assume in the following that conditions hold so that $\mathbb{H}^d_p(r)$ is a spacetime and we refer the reader to the appendix for details.

For any pair of points $x \subset y$ (i.e., $x \subset y$) and $y \subset y$ (i.e., $x \subset y$) and $y \subset x$ iff $(\chi_y, \chi_y) < 0$ and $\beta \leq -\varepsilon^2$. There exists no path between $x$ and $y$ (i.e., $x \subset y$ and $y \subset x$) iff $\beta \geq 0$.

**Lemma 3.1.** Assuming $\chi_y$ is timelike, $\chi_y$ is future-directed timelike (i.e., $\beta < 0$ and $(\chi_y, \chi_y) < 0$) and $\beta \leq -\varepsilon^2$. There exists no path between $x$ and $y$ (i.e., $x \subset y$ and $y \subset x$) iff $\beta \geq 0$.

- **The anti-de Sitter spacetime** $\mathbb{H}_p^d(r)$ and its projective version $\mathbb{H}_p^d(r) := \mathbb{H}_p^d(r) \pm 1$ are non-chronological and satisfy $x \subset y \implies y \subset x$, which is convenient to represent graphs with directed cycles. [Sim et al. 2021] use $\mathbb{H}_p^d(r)$ to represent directed graphs but also promote arcs (i.e., causal relation) between pairs of nodes that are not connected by any geodesic, which makes their problem hard to optimize. In this work, we only consider the existence of arcs if there exists a timelike geodesic in the convex normal neighborhood $V_x$ joining two events. See Appendix [C] for details.

- **A warped product** is a manifold $(M_1 \times f g_2)$ denoted by $M_1 \times f g_2$ where $(M_1, g_1)$, $(M_2, g_2)$ and $f : M_1 \to (0, \infty)$ is a smooth function called the warping function (e.g., $f = 1$). Let $M = B \times f$ be a Lorentz warped product where $B$ is Lorentz and $f$ is a complete Riemannian manifold. $M$ is time-orientable iff $B$ is (see page 417 of [O'Neill 1983]). $M$ satisfies the chronology, causality or strong causality iff $B$ does. Our framework can then be extended to warped products.

### 3.3 Lorentzian distance and Lorentzian length spaces

The Lorentzian distance indicates chronological order (hence causality) between events when it is positive, satisfies the reverse triangle inequality and its squared function is of class $C^2$ on a normal neighborhood. These properties make it ideal for optimization, and it can be used as time separation function. We refer the reader to Chapter 4 of [Beem et al. 1996] and Section 5 of [Minguzzi 2019].

**Proposition 4.5.3 of Hawking & Ellis (1973):** Let $x$ and $y$ lie in a convex normal neighborhood $V_x \subset U_x$. If $x$ and $y$ can be joined by a causal curve in $V_x$, the longest such curve is the unique causal geodesic in $V_x$ from $x$ to $y$. This is in contrast with Riemannian geometry where the (spacelike) geodesic corresponds to the shortest curve joining points. Since $V_x$ is a convex normal neighborhood, we can define the arc length of such a curve by $\chi_y(x, y) := \sqrt{-\langle \chi_y, \chi_y \rangle} \geq 0$. If $x \subset y$, $\chi_y(x, y)$ is called the Lorentzian distance from $x$ to $y$ on $V_x$, it corresponds to the elapsed proper time between the events $x$ and $y$ (i.e., as measured by a clock along the geodesic $\gamma_{\mathbb{R}^d}(t)$) [Gourgoulhon 2016]. If we define $\tau := \chi_y$, then $\langle V_x, d, \leq, \ll, \tau \rangle$ is called a Lorentzian length space. In this paper, we consider the squared Lorentzian distance which is defined as $y \in \mathcal{I}^+(x, V_x) \mapsto \chi_y^2(x, y) = -\langle \chi_y, \chi_y \rangle$. 


When evaluating learned representations (i.e., at test time), we therefore consider that \( \chi^2 \) without cycles. We propose to consider the chronological future and past that are intrinsic to our approach (see Appendix F) to learn our representations. Our approach could be extended to time-oriented pseudo-Riemannian manifolds and have been studied for decades to define the causal structure. For instance, it is known that causal relations induce partially ordered sets (called causal sets) on causal spacetimes and can for instance be generalized to timelike tangent vectors \( V \). If there exists a future-directed timelike curve joining them. Lemma 3.1 that exploits parallel transport \( \nu \), \( R \) functions also becomes straightforward and we use their sign to determine the direction of an edge. The negative of the pseudo-Riemannian gradient is generally not a descent direction. We then use the optimization tools introduced in Gao et al. (2018); Law & Stam (2020); Law (2021) as described in Section 4.3.

### 4 Related Work

Exploiting spacetimes to represent directed graphs was proposed by Clough & Evans (2017) to extend MultiDimensional Scaling (MDS) (Kruskal, 1964) to the simplest spacetime \( \mathbb{R}^d \). As in standard MDS, a target distance matrix between nodes is given as input and the goal of the method is to find the vector representations of nodes that return the best approximation of the target distance matrix when using the squared Lorentzian distance as the dissimilarity function. However, as discussed in Section 3.3, it is difficult to define target distances between pairs of nodes that are not connected. Moreover, the formulation in (Clough & Evans, 2017) considers only \( \mathbb{R}^d \) and does not account for future or past-direction between pairs of nodes during training. We solve this problem in Section 3.2 by enforcing future timelike geodesics to have their length smaller than some threshold \( \epsilon \in (0, \infty] \).

Sim et al. (2021) extended Clough & Evans (2017) to the anti-de Sitter space and Lorentz cylinder. Although our motivation is similar, our contributions are methodological, rely on the intuitions of Lorentzian pre-length spaces, and provide an easier interpretation of the learned representations as we explain below. First, Sim et al. (2021) do not address clearly the case when there is no geodesic between pairs of points, and their optimization framework leads to a distance loss term with a zero gradient in this case, which is difficult to optimize. Moreover, their prediction of an arc between a pair of nodes is determined via a Triple Fermi-Dirac (TFD) probability function that accounts for the distance between the nodes, the time coordinate difference \( \Delta t \) and its opposite value \( -\Delta t \) (i.e., TFD accounts for both the chronological future and past of a given node). In contrast, we restrict the representation of nodes connected by an arc to belong to \( \mathbb{R}^d \) (see pages 240-242 of O’Neill, 1983). We also have \( \mathbb{R}^d = \mathbb{R}^d = V \) if \( \mathbb{R}^d = \mathbb{R}^d = V \) is globally hyperbolic. Otherwise, we do not formulate our time separation function in the same way and we create arcs differently. Further comparisons with Sim et al. (2021) are provided in Appendix E.

The negative of the pseudo-Riemannian gradient is generally not a descent direction. We then use the optimization tools introduced in Gao et al. (2018); Law & Stam (2020); Law (2021) as described in Appendix E to learn our representations. Our approach could be extended to time-oriented pseudo-Riemannian manifolds of index \( \nu > 1 \) (see pages 240-242 of O’Neill, 1983). We also have \( x \ll y \) if there exists a future-directed timelike curve joining them. Lemma 3.1 that exploits parallel transport can for instance be generalized to timelike tangent vectors \( t_1, \ldots, t_{\nu} \) to define different types of arcs.

Some Riemannian approaches (Bordes et al., 2013; Ganea et al., 2018; Ylinis et al., 2018) have been used to represent DAGs. Ganea et al. (2018) consider entailment relations where \( x \ll y \) means that \( y \) is a subconcept of \( x \) by constructing cones for each node representation in hyperbolic geometry. The DAGs they consider are directed trees and their underlying undirected graphs are trees (i.e., graphs without cycles). We propose to consider the chronological future and past that are intrinsic to our manifolds and have been studied for decades to define the causal structure. For instance, it is known that causal relations induce partially ordered sets (called causal sets) on causal spacetimes and can then represent DAGs (Bombelli et al., 1987). Our framework can represent DAGs that are not directed trees (see example in Appendix G.1) and is not limited to DAGs (see examples in Appendix G.2).
Table 1: Link prediction for directed graphs. Median average precision (AP) percentages across 20 random initializations on a held-out test set.

| Dataset                  | $d$ | 3   | 5   | 10  | 50  | 100 | 3   | 5   | 10  | 50  | 100 |
|--------------------------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| Euclidean + FD           | 33.0| 34.2| 40.2| 44.5| 49.0| 29.4| 32.9| 39.7| 39.8| 34.8|
| Hyperbolid + FD          | 29.2| 37.9| 46.5| 48.8| 47.9| 28.8| 46.8| 50.8| 50.9| 52.5|
| Minkowski + TFD          | 34.7| 38.6| 44.9| 47.5| 49.4| 38.1| 45.2| 51.9| 55.6| 56.0|
| Anti de-Sitter + TFD     | 37.2| 41.3| 44.9| 47.5| 49.4| 38.1| 45.2| 51.9| 55.6| 56.0|
| Cylindrical Minkowski + TFD | 37.4| 42.7| 46.8| 53.4| 54.6| 41.0| 48.4| 56.3| 58.9| 61.0|

5 EXPERIMENTS

We show how our framework can represent graphs with directed cycles and predict links effectively in Section 5.1. We also show how the causal interpretation of our model can be used to represent hierarchical graphs with cycles in Section 5.2. We provide experiments on DAGs in Appendix D.1.

5.1 LINK PREDICTION ON GRAPHS WITH DIRECTED CYCLES

We now consider the link prediction task on the *Saccharomyces cerevisiae*, in silico and *Escherichia coli* DREAM5 datasets (Marbach et al. 2012) (see values of hyperparameters, results on *Escherichia coli* with standard deviations and more discussion in Appendix D.2). We follow the experimental protocol of Sim et al. (2021) by considering only the positive-regulatory nodes from the networks mentioned above while omitting the gene-expression data itself. Each network is randomly split into train and test sets, following 85/15 splits, and a part of the training set is used for validation.

Sim et al. (2021) use the Minkowski space $\mathbb{R}^d$, the anti de-Sitter space $\mathbb{H}^d(r)$, and the Cylindrical Minkowski space $\mathbb{L}^d(C)$. Both $\mathbb{H}^d(r)$ and $\mathbb{L}^d(C)$ are non-chronological and can then represent graphs with directed cycles. Sim et al. (2021) design a probability function that they call *Triple Fermi-Dirac (TFD)* specifically for their Cylindrical Minkowski space.

Instead, we propose to define the chronological future $\mathcal{T}^+(x, \mathcal{V}_x)$ of $x \in \mathbb{L}^d(C)$ such that if $\vec{x} \vec{y}$ is timelike, we have $y \in \mathcal{T}^+(x, \mathcal{V}_x)$ if $\exists k \in \mathbb{Z}, y_0 + kC \in (x_0, x_0 + C/2)$. Similarly, the chronological past $\mathcal{T}^-(x, \mathcal{V}_x)$ is defined such that if $\vec{x} \vec{y}$ is timelike, we have $y \in \mathcal{T}^-(x, \mathcal{V}_x)$ if $\exists k \in \mathbb{Z}, y_0 + kC \in (x_0 - C/2, x_0)$. In detail, we define the time separation function for $\mathbb{L}^d(C)$ as follows:

$$\tau(x, y) := \left(\left(\left(y_0 - x_0 + \frac{C}{2}\right) \mod C - \frac{C}{2}\right) \in \left[-\frac{C}{2}, \frac{C}{2}\right]\right)$$

where we use the modulo operation for real values which can be written as follows: $a \mod b := a - b \cdot ⌊a/b⌋$, and $⌊·⌋$ is the floor function.

Following our spacetime interpretation, we propose to learn our representations by optimizing:

$$\min_{(x_i \in \mathcal{M})_{i=1}^n} - \sum_{(x_i, x_j) \in E} \log(F(x_i, x_j)) - \sum_{(x_i, x_j) \not\in E} \log(1 - F(x_i, x_j))$$

(6)

where $F(x_i, x_j) := \sigma_{\theta_1}(\chi_{\theta_2}^m(x_i, x_j)) \cdot \sigma_{\theta_2}(\tau(x_i, x_j))$, $\theta_1, \theta_2 > 0$ are temperature hyperparameters, $m > 0$ is an exponent, and $\sigma_{\theta}(x) := 1/(1 + e^{-x/\theta})$ is the sigmoid function. We refer to Section C.2 and C.3 for the formulations of the time separation function $\tau$ and the squared Lorentzian distance $\chi_{\theta_2}$, respectively. This promotes future-directed timelike geodesics only between nodes connected by an arc. We report in Table 1 the Average Precision scores of our method (i.e., using equation 6) that outperforms baselines reported in Sim et al. (2021). The non-chronological cylindrical Minkowski space obtains a higher performance gap in low-dimensional space. This suggests it accounts for directed cycles better than chronological spacetimes. On the other hand, the gap is reduced in higher dimension except on *Escherichia coli* where the gap increases (see Table 4).

5.2 HIERARCHY EXTRACTION ON A SOCIAL NETWORK DATASET

Pseudo-Riemannian manifolds were introduced in the machine learning literature to extract hierarchies in graphs with cycles (Law & Stanić 2020). Spacetimes are a special family of pseudo-Riemannian manifolds whose time-orientation was not taken into account in Law & Stanić (2020). In this paper, we found that spacetimes are able to consistently outperform other pseudo-Riemannian manifold approaches by exploiting the causality interpretation implicit in the squared Lorentzian distance.
where the lower the metric, the better. ↑ the larger the metric (in absolute value), the better.

<table>
<thead>
<tr>
<th>Manifold</th>
<th>Problem</th>
<th>d(x, y)</th>
<th>Rank of 1st leader (↓)</th>
<th>Rank of 2nd leader (↓)</th>
<th>Top 5 ρ (↑)</th>
<th>Top 10 ρ (↑)</th>
</tr>
</thead>
<tbody>
<tr>
<td>R² (Euclidean)</td>
<td>equation 7</td>
<td>dᵢ(x, y)</td>
<td>11.4 ± 4.3</td>
<td>14.0 ± 2.4</td>
<td>-0.17 ± 0.70</td>
<td>-0.19 ± 0.40</td>
</tr>
<tr>
<td>P²(a) (hyperbolic)</td>
<td>equation 7</td>
<td>dᵢ(x, y)</td>
<td>7.0 ± 1.4</td>
<td>8.8 ± 1.1</td>
<td>-0.47 ± 0.31</td>
<td>-0.26 ± 0.13</td>
</tr>
<tr>
<td>P²(b) (elliptic)</td>
<td>equation 7</td>
<td>dᵢ(x, y)</td>
<td>7.5 ± 0.5</td>
<td>9.1 ± 1.2</td>
<td>-0.59 ± 0.03</td>
<td>-0.38 ± 0.08</td>
</tr>
<tr>
<td>P²(c)</td>
<td>equation 7</td>
<td>dᵢ(x, y)</td>
<td>2.2 ± 1.1</td>
<td>6.4 ± 2.2</td>
<td>0.28 ± 0.26</td>
<td>0.55 ± 0.20</td>
</tr>
<tr>
<td>P²(d)</td>
<td>equation 7</td>
<td></td>
<td>2.2 ± 1.1</td>
<td>6.4 ± 2.2</td>
<td>0.28 ± 0.26</td>
<td>0.55 ± 0.20</td>
</tr>
</tbody>
</table>

Graph dataset. We consider an undirected graph G = (V, E) where V = {vi}i=1 is the node set and E is the edge set such that (vi, vj) ∈ E indicates that vi and vj are connected by an edge. Zachary’s karate club dataset (Zachary[1977]) is a social network of n = 34 members of the karate club, each represented by a node vi. The club was split in two factions due to a conflict between the instructor (node v1) and the administrator (node v34) that are the two most important members (i.e., leaders) of the dataset. The other members had to decide between joining the new club created by v1, or stay with v34. Two members are joined by an edge (vi, vj) ∈ E if the members are friends.

In the original ultrahyperbolic approach [Law & Stam [2020] Law [2021]], each node vi is represented by a point xi ∈ M on some pseudo-Riemannian manifold M. The goal is to learn embeddings \{xi\}i=1 so that pairs of nodes joined by an edge (vi, vj) ∈ E are closer to each other than pairs of nodes not joined by an edge. To this end, the embeddings are learned in [Law & Stam [2020], Law [2021]] by minimizing the following problem:

\[
\min_{\{x_i \in M\}^n_{i=1}} - \sum_{(v_i,v_j) \in E} \log s - \frac{e^{d(x_i,x_j)/\theta}}{s} + \sum_{(u,v) \notin E} \frac{e^{-d(x_i,x_k)/\theta}}{s}
\]

where the dissimilarity function d(xi, xj) is the arc length d(xi, xj) := √(\|x_i^2 - x_j^2\|) of the geodesic joining x_i and x_j. If M is Riemannian, d_i is the Riemannian distance. In this paper, we propose instead to consider in equation [7] that d is the squared Lorentzian distance \^2E, as defined in Section [3.3] where \^E is the (convex) maximal normal neighborhood of x ∈ M.

We also propose to simply enforce the embeddings of vi and vj to be joined by a timelike geodesic if (vi, vj) ∈ E. Since G is an undirected graph, the future- or past-direction is not provided so we do not constrain the future-direction of the geodesics during training. Our problem formulation is then:

\[
\min_{\{x_i \in M\}^n_{i=1}} \sum_{(v_i,v_j) \in E} \sigma_i d(x_i,x_j) + \lambda \sum_{(v_i,v_j) \notin E} \sigma_i (-d(x_i,x_j))
\]

where \(\lambda > 0\) is a regularization parameter and we set d(xi, xj) = \^2E(xi, xj). The main goal of equation [8] is to enforce pairs of nodes (vi, vj) ∈ E to be joined by a timelike geodesic and pairs (vi, vj) \notin E to be joined by a spacelike geodesic since there is no causality between them.
We have proposed a general framework to represent data, and in particular nodes of a directed graph, in a spacetime. Compared to previous work, our framework is properly endowed with the structure of a Lorentzian pre-length space, which makes it easy to optimize and applicable to a large family of spacetimes. In the case of hierarchical undirected graphs, we show that we can enforce causality between nodes connected by an edge to extract the most important nodes in the hierarchy.
ACKNOWLEDGMENTS

We thank Yuval Atzmon, Gal Chechik and Haggai Maron for helpful discussions during the initial phase of this project. We also thank Aaron Sim for sharing his source code, Rafid Mahmood, Jos Stam and the anonymous reviewers for valuable feedback on early versions of the manuscript.

REFERENCES


A SUMMARY OF SUPPLEMENTARY MATERIAL

The supplementary material is structured as follows:
- Section B gives the differential geometry tools that are general to pseudo-Riemannian manifolds of constant curvature (called space forms).
- Section C gives the differential geometry tools that are specific to our spacetimes such as the time separation function and the squared Lorentzian distance.
- Section D presents additional results and further experimental details.
- Section E is an extended related work section.
- Section F gives the details of the optimizers we use.
- Section G gives examples of directed graphs described with our model.

B DIFFERENTIAL GEOMETRY OF PSEUDO-RIEMANNIAN SPACE FORMS

We provide the necessary differential geometry tools to work on the pseudo-Euclidean space $\mathbb{R}^d$, the pseudo-sphere $S^d(r)$ and the pseudo-hyperboloid $H^d(r)$. Most of them are explained in (Law, 2021; Law & Stam, 2020). We refer the reader to (Hawking & Ellis, 1973; O’Neill, 1983; Wolf, 2011).

B.1 PSEUDO-EUCLIDEAN SPACE

We recall that $\forall u, v \in \mathbb{R}^d, (u, v)_\nu = u^T \nu u$. Its geodesic $\gamma_{x \rightarrow y} : \mathbb{R} \rightarrow \mathbb{R}^d$ is $\gamma_{x \rightarrow y}(t) := x + ty$. The exponential map at $x$ is defined as $\exp_x(y) := \gamma_{x \rightarrow y}(1) = x + y$ and its inverse is $\gamma_{x \rightarrow y}^\prime(1) = \exp_x^{-1}(y) = y - x$. Moreover, $\forall x \in \mathbb{R}^d, y \in \mathbb{R}^d$, the parallel transport $\Gamma_x : T_x \mathbb{R}^d \rightarrow T_y \mathbb{R}^d$ can be defined as the identity function due to the isomorphism $T_x \mathbb{R}^d \cong T_y \mathbb{R}^d$. See page 67 of O’Neill (1983).

B.2 PSEUDO-SPHERE $S^d(r)$

We give here the differential geometry tools specific to the pseudo-sphere which is defined as the following hypersurface: $S^d(r) := \{ x \in \mathbb{R}^d : (x, x)_\nu = r^2 \}$. The tangent space $T_x S^d(r)$ of $S^d(r)$ at $x$ can be defined as: $T_x S^d(r) := \{ u \in \mathbb{R}^{d+1} : (u, x)_\nu = 0 \}$. In the case of the pseudo-sphere, we have $\forall u \in T_x S^d(r), \forall v \in T_x S^d(r), g_x(u, v) = (u, v)_\nu$.

Geodesic. The geodesic $\gamma_{x \rightarrow u} : \mathbb{R} \rightarrow S^d(r)$ satisfying $\gamma_{x \rightarrow u}(0) = x$ and $\gamma_{x \rightarrow u}^\prime(0) = u \in T_x S^d(r)$ is formulated for all $t \in \mathbb{R}$:

$$
\gamma_{x \rightarrow u}(t) := \begin{cases} 
\cos \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) x + \frac{r}{\sqrt{(u, u)_\nu}} \sin \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) u & \text{if } (u, u)_\nu > 0 \\
x + tu & \text{if } (u, u)_\nu = 0 \\
\cosh \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) x + \frac{r}{\sqrt{(u, u)_\nu}} \sinh \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) u & \text{if } (u, u)_\nu < 0 
\end{cases}
$$

Exponential map. The exponential map $\exp_x : T_x S^d(r) \rightarrow S^d(r)$ is defined such that $\forall u \in T_x S^d(r), \exp_x(u) = \gamma_{x \rightarrow u}(1)$. We then have:

$$
\exp_x(u) := \begin{cases} 
\cos \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) x + \frac{r}{\sqrt{(u, u)_\nu}} \sin \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) u & \text{if } (u, u)_\nu > 0 \\
x + u & \text{if } (u, u)_\nu = 0 \\
\cosh \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) x + \frac{r}{\sqrt{(u, u)_\nu}} \sinh \left( \frac{\sqrt{(u, u)_\nu}}{r} \right) u & \text{if } (u, u)_\nu < 0 
\end{cases}
$$

Logarithmic map. The logarithmic map $\log_x$ is the inverse of the exponential map $\exp_x$ on a normal neighborhood of $x \in S^d(r)$ denoted by $U_x = \{ y \in S^d(r) : \frac{(x, y)_\nu}{r^2} > -1 \}$. It is formulated:

$$
\forall y \in U_x, \frac{y-x}{\sinh(t)} := \log_x(y) := \begin{cases} 
\arccos \left( \frac{(x, y)_\nu}{r^2} \right) \left( y - \frac{(x, y)_\nu}{r^2} x \right) & \text{if } \frac{(x, y)_\nu}{r^2} \in (-1, 1) \\
y - x & \text{if } \frac{(x, y)_\nu}{r^2} = 1 \\
\arctanh \left( \frac{(x, y)_\nu}{r^2} \right) \left( y - \frac{(x, y)_\nu}{r^2} x \right) & \text{if } \frac{(x, y)_\nu}{r^2} > 1 
\end{cases}
$$
where \(\arccos\) is the inverse of the cosine function, and \(\arccosh\) is the inverse of the hyperbolic cosine function. One can verify from equation (10) and equation (11) that \(\bar{x} \cdot \bar{y}\) is timelike (i.e., \(<\bar{x} \cdot \bar{y}, \bar{x} \cdot \bar{y}> < 0\) iff \(<x, y>_{\nu} > r^2\).

Geodesic “distance”. As explained in (Law & Stam [2020]) and Chapter 5 of (O’Neill [1983]), when the logarithmic map exists for some pseudo-Riemannian manifold \(M\), the arc length of the geodesic \(\gamma_{x \rightarrow \bar{x}}\) from \(x \in M\) to \(y \in M\) corresponds to the radius function: \(\sqrt{|g_x(\bar{x} \cdot \bar{y}, \bar{x} \cdot \bar{y})|}\) where \(g_x : T_xM \times T_xM \rightarrow \mathbb{R}\) is the metric tensor at \(x\) and \(\bar{x} \cdot \bar{y} = \log_x(y)\) is the logarithmic map. The geodesic distance \(d_{\gamma} : \mathcal{S}_{\nu}^d(r) \times \mathcal{S}_{\nu}^d(r) \rightarrow \mathbb{R}\) is then:

\[
d_{\gamma}(x, y) := \sqrt{|\langle x \cdot y, x \cdot y \rangle_{\nu}|} = \begin{cases} \displaystyle r \arccosh \left( \frac{|x \cdot y|}{r} \right) & \text{if } |x \cdot y|_{\nu} \geq 1 \\ \displaystyle r \arccos \left( \frac{|x \cdot y|}{r^2} \right) & \text{if } |x \cdot y|_{\nu} \in (1, 1) \end{cases}
\]

(12)

\(d_{\gamma}\) is not a “distance metric” but a symmetric premetric: it satisfies (i) \(d_{\gamma}(x, y) = d_{\gamma}(y, x) \geq 0\) and (ii) \(d_{\gamma}(x, x) = 0\).

In (O’Neill [1983]), the “minimizing geodesic” is defined by its arc length and then also corresponds to the geodesic distance mentioned above.

Given the minimizing geodesic \(\gamma\) connecting \(x\) to \(y\), the parallel transport \(\Gamma_{\nu}^x : T_x \mathcal{S}_{\nu}^d(r) \rightarrow T_y \mathcal{S}_{\nu}^d(r)\) is a linear isometry such that \(\forall u, v, \langle u, v \rangle_{\nu} = \langle \Gamma_{\nu}^x(u), \Gamma_{\nu}^x(v) \rangle_{\nu}\). The parallel transport along \(\gamma\) from \(x = \gamma(0)\) to \(y = \gamma(1)\) (where \(x\) and \(y\) satisfy \(\langle x, y \rangle_{\nu} > -r^2\)) is:

\[
\Gamma_{\nu}^x(u) := u - \frac{\langle y, u \rangle_{\nu}}{|y, y|_{\nu} + r^2} (y + x)
\]

(13)

Theorem B.1 (Diffeomorphism (Wolf [2011]). There exists a diffeomorphism \(\psi : \mathcal{S}_{\nu}^d(r) \rightarrow \mathbb{R}^d - \nu\) \(\times \mathcal{S}_{\nu}^d(0)\). Let us note \(x = \left( \begin{array}{c} t \\ s \end{array} \right) \in \mathcal{S}_{\nu}^d(r)\) with \(t \in \mathbb{R}^d\) and \(s \in \mathbb{R}^{d-\nu+1}\). Let us note \(z = \left( \begin{array}{c} t \\ v \end{array} \right) \in \mathbb{R}^d \times \mathcal{S}_{\nu}^{d-\nu}(r)\) where \(v \in \mathcal{S}_{\nu}^{d-\nu}(r)\). The mapping \(\psi\) and its inverse \(\psi^{-1}\) can be formulated:

\[
\psi(x) = \left( \begin{array}{c} t \\ \frac{r}{\|x\|} s \end{array} \right) \text{ and } \psi^{-1}(z) = \left( \begin{array}{c} t \\ \frac{r}{\sqrt{r^2 + |x|^2}} v \end{array} \right).
\]

(14)

where \(\|\cdot\|\) is the standard Euclidean norm (i.e., \(\|s\| := \sqrt{s^t s}\)).

B.3 Pseudo-hyperboloid \(\mathbb{H}_{\nu}^d(r)\)

We recall the differential geometry tools (from (Law & Stam [2020])) specific to the pseudo-hyperboloid which is defined as the set: \(\mathbb{H}_{\nu}^d(r) := \{ x \in \mathbb{R}^{d+1}_+ : \langle x, x \rangle_{\nu+1} = -r^2 \}\). The tangent space \(T_x \mathbb{H}_{\nu}^d(r)\) at \(x\) can be defined as: \(T_x \mathbb{H}_{\nu}^d(r) := \{ u \in \mathbb{R}^{d+1}_+ : \langle u, x \rangle_{\nu+1} = 0 \}\). In the case of \(\mathbb{H}_{\nu}^d(r)\), we have \(\forall u \in T_x \mathbb{H}_{\nu}^d(r), \forall v \in T_x \mathbb{H}_{\nu}^d(r), g_x(u, v) = \langle u, v \rangle_{\nu+1}\).

Geodesic. The geodesic \(\gamma_{\bar{x} \rightarrow \bar{u}} : \mathbb{R} \rightarrow \mathbb{H}_{\nu}^d(r)\) satisfying \(\gamma_{\bar{x} \rightarrow \bar{u}}(0) = x\) and \(\gamma_{\bar{x} \rightarrow \bar{u}}(1) = u\) in \(T_x \mathbb{H}_{\nu}^d(r)\) is formulated for all \(t \in \mathbb{R}\):

\[
\gamma_{\bar{x} \rightarrow \bar{u}}(t) = \begin{cases} \cos \left( \frac{t \sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) x + \frac{r}{\sqrt{|\langle u, u \rangle_{\nu+1}|}} \sin \left( \frac{t \sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) u & \text{if } |\langle u, u \rangle_{\nu+1}| < 0 \\ \cosh \left( \frac{t \sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) x + \frac{r}{\sqrt{|\langle u, u \rangle_{\nu+1}|}} \sinh \left( \frac{t \sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) u & \text{if } |\langle u, u \rangle_{\nu+1}| > 0 \end{cases}
\]

(15)

Exponential map. The exponential map \(\exp_{\bar{x}} : T_x \mathbb{H}_{\nu}^d(r) \rightarrow \mathbb{H}_{\nu}^d(r)\) is defined such that \(\forall u \in T_x \mathbb{H}_{\nu}^d(r), \exp_{\bar{x}}(u) = \gamma_{\bar{x} \rightarrow \bar{u}}(1)\). We then have:

\[
\exp_{\bar{x}}(u) = \begin{cases} \cos \left( \frac{\sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) x + \frac{r}{\sqrt{|\langle u, u \rangle_{\nu+1}|}} \sin \left( \frac{\sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) u & \text{if } |\langle u, u \rangle_{\nu+1}| < 0 \\ \cosh \left( \frac{\sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) x + \frac{r}{\sqrt{|\langle u, u \rangle_{\nu+1}|}} \sinh \left( \frac{\sqrt{|\langle u, u \rangle_{\nu+1}|}}{r} \right) u & \text{if } |\langle u, u \rangle_{\nu+1}| > 0 \end{cases}
\]

(16)
Logarithmic map. The logarithmic map log\(x\) is defined as the inverse of the exponential map exp\(x\) on a normal neighborhood of \(x \in \mathbb{H}_d^{d}(r)\) denoted by \(\mathcal{U}_x = \{y \in \mathbb{H}_d^{d}(r) : \frac{\langle x, y \rangle_{\nu+1}}{r^2} < 1\}\). It is formulated:

\[
\forall y \in \mathcal{U}_x, \quad \log_x(y) = \begin{cases} 
\frac{\text{arccosh} \left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right)}{\sqrt{\left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right)^2 - 1}} (y + \frac{\langle x, y \rangle_{\nu+1}}{r^2} x) & \text{if } \frac{\langle x, y \rangle_{\nu+1}}{r^2} < -1 \\
y - x & \text{if } \frac{\langle x, y \rangle_{\nu+1}}{r^2} = -1 \\
\frac{\text{arccos} \left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right)}{\sqrt{1 - \left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right)^2}} (y + \frac{\langle x, y \rangle_{\nu+1}}{r^2} x) & \text{if } \frac{\langle x, y \rangle_{\nu+1}}{r^2} \in (-1, 1)
\end{cases}
\]

(17)

One can verify that a nonconstant geodesic from \(x\) to \(y\) is timelike iff \(\langle x, y \rangle_{\nu} \in (-r^2, r^2)\) or \(y = \pm x\).

Geodesic “distance”. The geodesic distance \(d_\gamma : \mathbb{H}_d^{d}(r) \times \mathbb{H}_d^{d}(r) \to \mathbb{R}\) is then:

\[
d_\gamma(x, y) = \sqrt{\langle x\hat{y}, x\hat{y} \rangle_{\nu+1}} = \begin{cases} 
r \text{arccosh} \left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right) & \text{if } \frac{\langle x, y \rangle_{\nu+1}}{r^2} < -1 \\
r \text{arccos} \left( -\frac{\langle x, y \rangle_{\nu+1}}{r^2} \right) & \text{if } \frac{\langle x, y \rangle_{\nu+1}}{r^2} \in (-1, 1)
\end{cases}
\]

(18)

Parallel transport on \(\mathbb{H}_d^{d}(r)\). The parallel transport connecting \(x \in \mathbb{H}_d^{d}(r)\) to \(y \in \mathbb{H}_d^{d}(r)\) is:

\[
\Gamma^x_y(u) = u - \frac{\langle y, u \rangle_{\nu+1}}{\langle x, y \rangle_{\nu+1} - r^2} (y + x) \quad \text{where } \langle x, y \rangle_{\nu+1} < r^2
\]

(19)

Theorem B.2 (Diffeomorphism [Wolf, 2011]). There exists a diffeomorphism \(\psi : \mathbb{H}_d^{d}(r) \to \mathbb{S}_0(r) \times \mathbb{R}^{d-\nu}\). Let us note \(x = \left(\begin{array}{c} t \\ s \end{array}\right) \in \mathbb{H}_d^{d}(r)\) with \(t \in \mathbb{R}^{\nu+1}\) and \(s \in \mathbb{R}^{d-\nu}\). Let us note \(z = \left(\begin{array}{c} u \\ v \end{array}\right) \in \mathbb{S}_0(r) \times \mathbb{R}^{d-\nu}\) where \(u \in \mathbb{S}_0(r)\) and \(v \in \mathbb{R}^{d-\nu}\). The mapping \(\psi\) and its inverse \(\psi^{-1}\) can be formulated:

\[
\psi(x) = \left(\frac{r^2 + \|v\|^2}{r^2}, t\right) \quad \text{and} \quad \psi^{-1}(z) = \left(\frac{\sqrt{r^2 + \|v\|^2}}{r^2} u, v\right).
\]

(20)

The anti-de Sitter spacetime \(\mathbb{H}_d^{d}(r)\) is non-chronological and satisfies \(x \ll y \implies y \ll x\), which is convenient to represent graphs with directed cycles. It was used in [Sim et al., 2021] to represent directed graphs. Nonetheless, it is worth noting that the problem formulation of [Sim et al., 2021] for \(\mathbb{H}_d^{d}(r)\) also promotes arcs (i.e., causal relation) between pairs of nodes that are not connected by any geodesic, which makes their problem hard to optimize. Following general relativity, we only consider the existence of arcs if there exists a timelike geodesic in \(V_x\) joining two events \(x\) and \(y\). See Appendix C for details.

B.4 PROJECTIVE SPACE \(\mathbb{P}_d^{d}(r)\)

The manifold \(\mathbb{P}_d^{d}(r) := \mathbb{H}_d^{d}(r)/\pm 1\) used in [Law, 2021] is time-orientable for all \(d \geq 2\) (see page 214 of [O’Neill, 1983]). We refer the reader to [Law, 2021] for details about \(\mathbb{P}_d^{d}(r) := \mathbb{H}_d^{d}(r)/\pm 1\).

C FUTURE DIRECTION, TIME SEPARATION FUNCTION AND LORENTZIAN DISTANCE

C.1 FUTURE DIRECTION

We discuss in this section how to constrain future direction for our spacetimes.

C.1.1 MINKOWSKI SPACE

The explanation can be found in Section 3.2 when \(M = \mathbb{R}_1^{d}\). We recall that we defined: \(t := (1, 0, \ldots, 0)^T, \hat{x} = y - x, \langle \hat{x} \hat{y}, \hat{x} \hat{y} \rangle = \langle \hat{x} \hat{y}, \hat{y} \hat{x} \rangle_1\) and \(\alpha := -(y_0 - x_0) + \sqrt{\sum_{i=1}^{d-1} (y_i - x_i)^2}\).

If \(\alpha\) is negative, then \(\hat{x} \hat{y}\) is timelike (i.e., \(\langle \hat{x} \hat{y}, \hat{x} \hat{y} \rangle < 0\)) and \(\hat{x} \hat{y} \in C_+^+ (t)\). By definition, \(\hat{x} \hat{y}\) is then future-directed timelike.
C.1.2 De Sitter Space

We provide the proof of Lemma 3.1 which is written: when $\mathcal{M} = \mathbb{H}^d_1(r)$ and $\overline{x} \overline{y}$ is timelike, $\overline{x} \overline{y}$ is future-directed iff $\Gamma^+_p(t) \in C^+_x(\overline{x} \overline{y})$ if $\Gamma^+_p$ is defined (i.e., $p \in \mathcal{U}_\mathcal{X}$), and $\Gamma^-_{-p}(t) \in C^+_x(\overline{x} \overline{y})$ otherwise (i.e., $-p \in \mathcal{U}_\mathcal{X}$).

We first recall a property from page 146 of [O’Neill 1983]: If $\mathcal{M}$ is a Lorentz manifold, if a piecewise smooth curve $\alpha$ is timelike, this means not only that $\alpha'(t)$ is timelike, but at each break $t_i$ of $\alpha$:

$$\langle \alpha'(t^+), \alpha'(t^-) \rangle < 0.$$  \hspace{1cm} (21)

Here the first vector derives from $\alpha|_{[t_{i-1}, t_i]}$ and the second from $\alpha|_{[t_i, t_{i+1}]}$. Thus $\alpha'$ does not switch timecones at the break.

Let us note the poles $p = (0, \ldots, 0, r)^T \in S^d_1(r)$, $-p = (0, \ldots, 0, -r)^T \in S^d_1(r)$ and $t = (1, 0, \ldots, 0)^T$ where $t \in T_p S^d_1(r)$ and $t \in T_{-p} S^d_1(r)$. We also note $x := (x_0, x_1, \ldots, x_d)^T \in S^d_1(r)$.

If $x_d \in (-r, r)$, we show below that equation 21 can be satisfied by defining $\alpha$ such that $\alpha(0) = p$, $\alpha(t_1) = x$, $\alpha(t_2) = -p$, $\alpha'(0) = t$, $\alpha'(t_1) = \Gamma^+_p(t)$, $\alpha'(t_2) = \Gamma^-_{-p}(t)$, $\alpha'(t_2) = t$. For completeness, we also have $\forall t \in [t_0, t_1], \alpha'(t) = \Gamma^+_p(t)$ and $\forall t \in [t_1, t_2], \alpha'(t) = \Gamma^-_{-p}(t)$.

Let us arbitrarily consider that $t \in T_p S^d_1(r)$ defines the future direction of the manifold. For all $x \in S^d_1(r)$ that satisfies $\langle x, p \rangle > -r^2$ (i.e., $p \in \mathcal{U}_\mathcal{X}$), we note $\Gamma^+_x(t)$ the parallel translate of $t \in T_p S^d_1(r)$ to $T_x S^d_1(r)$. As explained in Appendix B.1.2 of [Law 2021], it is formulated:

$$\Gamma^+_p(t) := t - \frac{(x, t_1)}{(x, p)_1 + r^2} (p + x) = t + \frac{x_0}{r x_d + r^2} (p + x) \hspace{1cm} (22)$$

We know from page 66 of [O’Neill 1983] that parallel transport (also called parallel translation) is a linear isometry that satisfies:

$$\forall u \in T_x S^d_1(r), \forall v \in T_x S^d_1(r), \langle u, v \rangle = \langle \Gamma^+_x(u), \Gamma^+_x(v) \rangle = \langle \Gamma^+_x(u), \Gamma^+_x(v) \rangle \hspace{1cm} (23)$$

and $u = \Gamma^+_x(\Gamma^+_x(u))$. This implies

$$\Gamma^+_x(t) \in C^+_x(u) \iff \Gamma^+_x(u) \in C^+_x(t) \hspace{1cm} (24)$$

Similarly, if $\langle x, -p \rangle > -r^2$ (i.e., $-p \in \mathcal{U}_\mathcal{X}$), we have $\Gamma^-_{-p}(t) \in C^+_x(u) \iff \Gamma^-_{-p}(u) \in C^+_x(t)$.

Let us assume that $x$ satisfies both $p \in \mathcal{U}_\mathcal{X}$ (i.e., $\langle x, p \rangle > -r^2$) and $p \in \mathcal{U}_\mathcal{X}$ (i.e., $\langle x, p \rangle < r^2$), this is equivalent to $x$ satisfying $x_d \in (-r, r)$.

If $x_d \in (-r, r)$, then $\Gamma^+_x(p) \in C^+_x(u), \Gamma^-_{-p}(t) \in C^+_x(u) \iff \langle \Gamma^+_x(t), \Gamma^-_{-p}(t) \rangle < 0$. By definition, we have:

$$\Gamma^+_x(t) := t - \frac{(x, t_1)}{(x, -p)_1 + r^2} (-p + x) = t + \frac{x_0}{-r x_d + r^2} (-p + x) \hspace{1cm} (25)$$

$$\langle \Gamma^+_x(t), \Gamma^-_{-p}(t) \rangle = -1 - \frac{x_0^2}{r x_d + r^2} - \frac{x_0^2}{-r x_d + r^2} = -1 + 2 \frac{x_0^2}{(x_d + r)(-x_d + r)} < 0 \hspace{1cm} (26)$$

We have shown that $\Gamma^+_x(t)$ and $\Gamma^-_{-p}(t)$ have the same future direction by showing that $\langle \Gamma^+_x(t), \Gamma^-_{-p}(t) \rangle < 0$. A (piecewise) smooth curve that preserves causality can then be found by using $\Gamma^+_x(t)$ or $\Gamma^-_{-p}(t)$. \hfill $\square$

C.1.3 Anti-de Sitter Space

When $\mathcal{M} = \mathbb{H}^d_1(r)$, $\overline{x} \overline{y}$ is future-directed iff $\Gamma^+_p(t) \in C^+_x(\overline{x} \overline{y})$ if $\Gamma^+_p$ is defined (i.e., $p \in \mathcal{U}_\mathcal{X}$), and $\Gamma^-_{-p}(-t) \in C^+_x(\overline{x} \overline{y})$ otherwise (i.e., $-p \in \mathcal{U}_\mathcal{X}$). The proof is similar to the one in Section C.1.2.

Let us note the poles $p = (r, 0, \ldots, 0)^T \in \mathbb{H}^d_1(r)$, $-p = (-r, 0, \ldots, 0)^T \in \mathbb{H}^d_1(r)$ and $t = (0, 1, 0, \ldots, 0)^T$ where $t \in T_p \mathbb{H}^d_1(r)$ and $t \in T_{-p} \mathbb{H}^d_1(r)$. Let $x = (x_{-1}, x_0, \ldots, x_{d-1})^T \in \mathbb{H}^d_1(r)$. We recall that $p \in \mathcal{U}_\mathcal{X}$ if $\langle x, p \rangle_r < r^2$ (i.e., $x_{-1} > -r$) and $-p \in \mathcal{U}_\mathcal{X}$ if $\langle x, -p \rangle_r < r^2$ (i.e., $x_{-1} < r$). We define $x$ such that $x_{-1} \in (-r, r)$ (i.e., $p \in \mathcal{U}_\mathcal{X}$ and $-p \in \mathcal{U}_\mathcal{X}$).
Let us note $\Gamma^x_p(t)$ the parallel translate of $t \in T_p \mathbb{H}^d_1(r)$ to $T_x \mathbb{H}^d_1(r)$. As explained in Appendix B.3 of [Law 2021], it is formulated:

$$
\Gamma^x_p(t) := t - \frac{\langle x, t \rangle_2}{\langle x, p \rangle_2 - r^2}(p + x) = t + \frac{x_0}{-r x_{-1} - r^2}(p + x)
$$

We show below that $\Gamma^x_p(t)$ and $\Gamma^x_{-p}(-t)$ have the same future direction by showing that $\langle \Gamma^x_p(t), \Gamma^x_{-p}(-t) \rangle_2 < 0$.

$$
\Gamma^x_{-p}(-t) := -t - \frac{\langle x, -t \rangle_2}{\langle x, -p \rangle_2 - r^2}(-p + x) = -t - \frac{x_0}{-r x_{-1} - r^2}(-p + x)
$$

$$
\langle \Gamma^x_p(t), \Gamma^x_{-p}(-t) \rangle_2 = 1 + \frac{x_0^2}{r x_{-1} - r^2} + \frac{x_0^2}{-r x_{-1} - r^2} = 1 - 2 \frac{x_0^2}{r^2 - x_{-1}^2}
$$

$$
= \frac{r^2 - x^2_{-1} - 2 x_0^2}{r^2 - x^2_{-1}} < 0.
$$

Equation (30) is negative because we have by definition of $\mathbb{H}^d_1(r)$: $-x^2_{-1} - x^2_0 + \sum_{i=1}^{d-1} x_i^2 = -r^2$, which implies $x^2_0 \geq r^2 - x^2_{-1} > 0$, and the denominator is positive because we defined $x_{-1} \in (-r, r)$. □

C.2 TIME SEPARATION FUNCTION

We give the formulations we use for the time separation function $\tau(x, y)$ that is positive if $y$ is in the chronological future of $x$. There is no unique way to define it. We assume here that $\nu = 1$.

**Globally hyperbolic spacetimes.** If the spacetime $\mathcal{M}$ is globally hyperbolic, there exists a Cauchy time function $c : \mathcal{M} \rightarrow \mathbb{R}$ (i.e., for any $t \in \mathbb{R}$, the set $\Sigma_t = \{ y \in \mathcal{M} : c(y) = t \}$ is a Cauchy hypersurface) that can be used to define $\tau(x, y) := c(y) - c(x)$ since it satisfies the reverse triangle inequality $\tau(x, z) \geq \tau(x, y) + \tau(y, z)$ when $x \leq y \leq z$. With this formulation, it actually satisfies $\tau(x, z) = \tau(x, y) + \tau(y, z)$ when $x \leq y \leq z$.

In theory, to have a Lorentzian pre-length space, one could define $\tau$ such that it is 0 if $x \not\leq y$, but this is not easy to optimize in practice so we ignore this constraint for optimization purpose.

C.2.1 MINKOWSKI SPACE $\mathbb{R}^d_1$

If $\mathcal{M} = \mathbb{R}^d_1$ then we arbitrarily define:

$$
\tau(x, y) := -\langle x y, t \rangle = -\langle x y, t \rangle_1 = y_0 - x_0
$$

where $t := (1, 0, \ldots, 0)^T$ defines the future timecone $C^+_1(t)$. It is worth noting that the function $c : \mathbb{R}^d_1 \rightarrow \mathbb{R}$ defined such that $\forall x = (x_0, \ldots, x_{d-1})^T \in \mathbb{R}^d_1, c(x) := x_0$ is a Cauchy time function.

C.2.2 DE SITTER SPACE $\mathbb{S}^d_1(r)$

Following our discussion in Section C.1.2 and using the formulation of the parallel transport in equation (27) we can formulate the time function as:

$$
\tau(x, y) := \begin{cases} 
-\langle \Gamma^x_p(x y), t \rangle_\nu & \text{if } \langle x, p \rangle_\nu \geq 0 \\
-\langle \Gamma^x_{-p}(x y), t \rangle_\nu & \text{otherwise.}
\end{cases}
$$

(32)

where $\nu = 1$. One limitation of equation (32) is that it assumes that $x y$ is defined, which might not be the case (if $y \not\in U_x$).

From the discussion in Chapter 5.2 and Figure 16 (i) of [Hawking & Ellis 1973] that uses Cartesian coordinates to formulate space coordinates on the pseudo-sphere, the function $c : \mathbb{S}^d_1(r) \rightarrow \mathbb{R}$ defined such that $\forall x = (x_0, \ldots, x_d)^T, c(x) := x_0$ is a Cauchy time function. One can then also simply define:

$$
\tau(x, y) := y_0 - x_0.
$$

(33)

In practice, we found that using equation (33) returns better performance than equation (32) in the experiments of Section D.1. We then also use it for the experiment of Section 5.1. See Section D.2 for details.
C.2.3 Anti-de Sitter space $\mathbb{H}_d^d(r)$

When $\nu > 0$, $\mathbb{H}_d^d(r)$ is non-chronological and satisfies $x \ll y \implies y \ll x$, which is convenient to represent graphs with directed cycles. There exists a future-directed timelike geodesic from $x$ to $y$ only if $\bar{xy}$ is timelike or $y = \pm x$.

Following our discussion in Section C.1.3 and using the formulation of the parallel transport in equation 27 and assuming equation 27, we can formulate the time function as:

$$\tau(x, y) := \begin{cases} -\langle \Gamma_p^x(\bar{xy}), t \rangle_{\nu+1} & \text{if } \langle x, p \rangle_{\nu+1} \leq 0 \\ -\langle \Gamma_p^x(\bar{xy}), -t \rangle_{\nu+1} & \text{otherwise.} \end{cases}$$

(34)

C.2.4 Projective version of the anti-de Sitter space $\mathbb{P}_d^d(r)$

We recall that each point of $\mathbb{P}_d^d(r)$ is an unordered pair $\{-x, x\} \in \mathbb{P}_d^d(r)$ where $x \in \mathbb{H}_d^d(r)$. By using the formulation of the parallel transport in equation 27 and assuming $\nu = 1$, we can then define:

$$\tau(\{-x, x\}, \{-y, y\}) := \begin{cases} -\langle \Gamma_p^x(\bar{xy}), t \rangle_{\nu+1} & \text{if } \langle x, p \rangle_{\nu+1} \leq 0 \\ -\langle \Gamma_p^x(\bar{xy}), -t \rangle_{\nu+1} & \text{otherwise.} \end{cases}$$

(35)

Let us assume that $p = (r, 0, \ldots, 0)^{\top}$. We can choose $x$ such that $x_{-1} \geq 0$ so that $\langle x, p \rangle_2 \leq 0$ is satisfied, and we can also choose $y$ so that $\langle x, y \rangle_2 \leq 0$ is satisfied. We can define $t := (0, 1, 0, \ldots, 0)^{\top} \in T_p \mathbb{P}_d^d(r)$. If we define $u = (u_{-1}, u_0, \ldots, u_{d-1})^{\top} = \Gamma_p^x(\bar{xy})$, then equation 35 can be rewritten:

$$\tau(\{-x, x\}, \{-y, y\}) = u_0$$

(36)

Another possibility is to use the following time separation function:

$$\tau(\{-x, x\}, \{-y, y\}) = u_0 - \sum_{i=1}^{d-1} u_i^2$$

(37)

which is positive only if $\bar{xy}$ is future-directed timelike.

C.2.5 Cylindrical Minkowski space $\mathbb{L}_d^d(C)$

As we explain in the main paper, we propose to define the chronological future $\mathcal{I}^+(x, V_x)$ of $x \in \mathbb{L}_d^d(C)$ such that if $\bar{xy}$ is timelike, we have $y \in \mathcal{I}^+(x, V_x)$ if $\exists k \in \mathbb{Z}, y_0 + kC \in (x_0, x_0 + C/2)$.

Similarly, the chronological past $\mathcal{I}^-(x, V_x)$ is defined such that if $\bar{xy}$ is timelike, we have $y \in \mathcal{I}^-(x, V_x)$ if $\exists k \in \mathbb{Z}, y_0 + kC \in (x_0 - C/2, x_0)$.

We define the time separation function for $\mathbb{L}_d^d(C)$ as follows:

$$\tau(x, y) := \left( (y_0 - x_0 + \frac{C}{2}) \mod \frac{C}{2} \right) - \frac{C}{2} \in [-\frac{C}{2}, \frac{C}{2}]$$

(38)

where we use the modulo operation for real values which can be written as follows: $a \mod b := a - b \left\lfloor \frac{a}{b} \right\rfloor$, and $\lfloor \cdot \rfloor$ is the floor function.

C.3 Squared Lorentzian distance

We give the formulation of the squared Lorentzian distance for the different spacetimes that we use in the main paper depending on the nature of $\mathcal{M}$:

If $\mathcal{M} = \mathbb{R}_1^d$, $\chi^2_d(x, y) := -\langle \bar{xy}, \bar{xy} \rangle = -\langle \bar{xy}, \bar{xy} \rangle_1 := (y_0 - x_0)^2 - \sum_{j=1}^{d-1} (y_j - x_j)^2$.

(39)
We explain here why our formulation of equation 4 is general and can be extended to other spacetimes. As explained in Theorem 2.7 of [Minguzzi 2019], any pair of its points $x, y$ that satisfy $x \ll y$ can be joined by a unique future-directed timelike geodesic $\gamma_{x \rightarrow x'y'}$, which is also the unique longest timelike curve joining $x$ to $y$. Since $\gamma_{x \rightarrow x'y'}$ is future-directed timelike, it satisfies $(\vec{x'y'}, \vec{x'y'}) < 0$ and $(\vec{x'y'}, t) < 0$ (i.e., $\vec{x'y'} \in C_{x}^{+}(t)$) where the timelike tangent vector $t \in T_{x} \mathcal{M}$ defines the future direction of $\mathcal{M}$. Moreover, the arc length of the geodesic $\gamma_{x \rightarrow x'y'}$ from $x$ to $y$ is $\sqrt{-\langle \vec{x'y'}, \vec{x'y'} \rangle}$, and is called its Lorentzian distance.

From Theorem 5.6 and Lemma 5.7 of [Minguzzi 2019], there exists a strongly causal open set containing $x$ such that its Lorentzian distance with any other point $y$ in that set has its Lorentzian distance $\sqrt{-\langle \vec{x'y'}, \vec{x'y'} \rangle}$ upper bounded by some constant $\varepsilon > 0$. We can then define that open set to be the convex normal neighborhood $\mathcal{V}_{x}$, and we can choose $\varepsilon > 0$ such that all the points $y \in \mathcal{V}_{x}$ that satisfy $x \ll y$ also satisfy $-\varepsilon^{2} < (\vec{x'y'}, \vec{x'y'}) < 0$.

Since $\mathcal{V}_{x}$ is a subset of the maximal normal neighborhood $\mathcal{U}_{x}$, we then obtain exactly the definition of equation 4 which is: $\mathcal{U}^{+}(x, \mathcal{V}_{x}) = \{ y \in \mathcal{U}_{x} : -\varepsilon^{2} < (\vec{x'y'}, \vec{x'y'}) < 0, \vec{x'y'} \in C_{x}^{+}(t) \}$.

In our experiments, we consider that $\mathcal{V}_{x} = \mathcal{U}_{x}$. When $\mathcal{M}$ is the Minkowski space $\mathbb{R}_{1}^{d}$ or the de Sitter space $\mathbb{S}_{1}^{d}(r)$, we then have $\varepsilon = +\infty$ but we describe some case where $\varepsilon$ might be finite in Section 3.2.
We optimize the problem:

$$\min_{(x_i \in \mathcal{M})_{n \geq 1}} \sum_{v_i \neq v_j \in E} \sigma_{\theta_1} \left( \chi^2_\theta(x_i, x_j) \right) + \lambda \sum_{v_i \neq v_j} \left( \sigma_{\theta_1} \left( -\chi^2_\theta(x_i, x_j) \right) + \sigma_{\theta_2} \left( -\tau \left( x_i, x_j \right) \right) \right)$$

(43)

where $$\sigma(y) := 1/(1 + e^{-y/\theta})$$ is the sigmoid function, $$\theta_1, \theta_2 > 0$$ are temperature parameters, $$\lambda$$ is a regularization parameter and $$\tau(x, y)$$ is a time separation function that is positive (resp. negative) if $$y$$ is in the chronological future (resp. past) of $$x$$. For instance, if $$\mathcal{M} = \mathbb{R}_+^d$$ then we arbitrarily define $$\tau(x, y) := -(\mathbf{x}^\top \mathbf{t}) = y_0 - x_0$$, where $$\mathbf{t} := (1, 0, \ldots, 0)^\top$$ defines the future timecone $$C^+(\mathbf{t})$$.

In this experiment, our temperature hyperparameter values are $$\theta_1 = \theta_2 = 1$$, and we fix the radius to $$r = 1$$. We run our experiments for $$10^8$$ iterations with a step size of $$10^{-6}$$ by using the optimization tools of [Law 2021, Law & Stam 2020]. The regularization parameter $$\lambda$$ is set to $$\lambda = \frac{|E|}{|E'|}$$ where $$|E|$$ is the number of pairs that satisfy $$v_i \ll v_j$$ and $$|E'|$$ is the number of pairs that satisfy $$v_a \nleq v_b$$.

We report in Table 3 how well the learned representations manage to preserve chronological order when we select the $$|V| = 200$$ or 1000 most cited papers. For instance if $$\mathcal{M} = \mathbb{R}_+^d$$, we report the percentage of pairs of nodes $$v_i \ll v_j$$ represented by $$x_i$$ and $$x_j$$ that satisfy $$(\mathbf{x}_i, \mathbf{x}_j, t) < 0$$. The chronological manifolds $$\mathbb{R}_+^d$$ and $$\mathbb{S}_1^d(r)$$ manage to predict chronological order better than the non-chronological manifold $$\mathbb{P}_1^d(r)$$. This suggests that chronological spacetimes are more appropriate to represent graphs that are almost DAGs. The time separation function in equation 33 returns better performance than equation 32, we then use it in the rest of our experiments.

Figure 3 illustrates the embeddings learned when $$\mathcal{M} = \mathbb{R}_+^d$$. Article representations tend to satisfy the chronological order along the time coordinate.

| $|V|$ | $\mathbb{R}_+^d$ | $\mathbb{R}_+^d$ | $\mathbb{R}_+^d$ | $\mathbb{P}_1^d(r)$ | $\mathbb{P}_1^d(r)$ | $\mathbb{P}_1^d(r)$ |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|
| 200 | 93.7% | 94.0% | 94.2% | 73.3% | 73.7% | 74.9% |
| 1000 | 91.5% | 91.8% | 92.0% | 72.2% | 72.4% | 72.3% |

| $|V|$ | $\mathbb{S}_1^d(r)$ | $\mathbb{S}_1^d(r)$ | $\mathbb{S}_1^d(r)$ | $\mathbb{S}_1^d(r)$ | $\mathbb{S}_1^d(r)$ | $\mathbb{S}_1^d(r)$ |
|-----|----------------|----------------|----------------|----------------|----------------|----------------|
| 200 | 93.8% | 93.8% | 93.7% | 94.4% | 94.9% | 94.9% |
| 1000 | 89.1% | 89.2% | 89.5% | 92.8% | 93.4% | 93.6% |
Figure 3: 2-dimensional representations of the Arxiv High-energy physics theory (HEP-TH) citation network (Gehrke et al., 2003) when $|V| = 200$ and $\mathcal{M} = \mathbb{R}^2$. The first four characters of the labels correspond to the year of submission on Arxiv, and the following two characters are the month of publication. For instance, the label “199208027” corresponds to the article submitted in August 1992 available at: [https://arxiv.org/abs/hep-th/9208027](https://arxiv.org/abs/hep-th/9208027). Similarly, “199409089” is the article available at [https://arxiv.org/abs/hep-th/9409089](https://arxiv.org/abs/hep-th/9409089) and submitted in September 1994. The embeddings more or less follow the chronological order (along the time axis) in which the articles were submitted to Arxiv.
D.2 Directed link prediction

In this section, we provide details about the experiments in Section 5.1.

We report in Table 4 the link prediction scores on the *Saccharomyces cerevisiae, in silico* and *Escherichia coli* DREAM 5 datasets (Marbach et al., 2012). Following the evaluation protocol of (Sim et al., 2021), we report the median + standard deviation across 20 random initialization. We use the same training and test splits as (Sim et al., 2021).

We use the following hyperparameters on the DREAM 5 datasets to define equation 6:

- if $M = S_1^d(r)$, $r = 1$, $\theta_1 = 0.15$, exponent $m = 1$, $\theta_2 = 0.03$, learning rate $= 10^{-5}$, number of epochs $= 2000$.
- if $M = L_1^d(C)$, $C = 8$, $\theta_1 = 0.15$, $\theta_2 = 0.03$, exponent $m = 1$, learning rate $= 10^{-3}$, number of epochs $= 2000$.
- if $M = R_1^d$, $\theta_1 = 0.15$, $\theta_2 = 0.03$, exponent $m = 1$, learning rate $= 10^{-3}$, number of epochs $= 2000$.

When $d = 10, 50$ or 100, $S_1^d(r)$ is not time-oriented (see explanation in Section 3.2). We report the scores obtained in these dimensionalities to be fair with baselines.

We also reran the experiments of Minkowski + TFD on the Escherichia coli DREAM 5 dataset.

It is worth noting that the DREAM 5 datasets contain a relatively small number of cycles:

- The *Saccharomyces cerevisiae* DREAM 5 dataset contains 9 nodes that are part of at least one directed cycle (0.5% of 1,994 nodes) and 19 edges that are part of at least one directed cycle (0.5% of 3,940).
- The *Escherichia coli* DREAM 5 dataset contains 18 nodes that are part of at least one directed cycle (1.7% of 1,081 nodes) and 23 edges that are part of at least one directed cycle (1.1% of 2,066).
- The *in silico* DREAM 5 dataset contains 28 nodes that are part of at least one directed cycle (1.8% of 1,565 nodes) and 39 edges that are part of at least one directed cycle (1.0% of 4,012).

The choice of a specific manifold acts as some inductive bias. When the manifold is chosen to be chronological, the created graph does not necessarily contain directed cycles but allows their existence. On the other hand, chronological manifolds are more appropriate for Directed Acyclic Graphs as they ensure that the created graph does not contain directed cycles. From our results, it seems that the (nonchronological) Cylindrical Minkowski spacetime obtains much better performance in the low-dimensional case as its lack of causality allows some flexibility that is less beneficial in the high-dimensional case.

Sim et al. (2021) also use the synthetic “Duplication-Divergence” (Dupdiv) dataset (Ispolatov et al., 2005), but their dataset contains only 100 edges and 1,026 edges. We generated a bigger version of Dupdiv that contains 1,000 edges and 26,649 edges (22,651 for training/validation and 3,998 for test) following the same setup. More precisely, 748 nodes (74.8%) are part of at least one directed cycle and 22,409 edges are part of at least one directed cycle (84.1% of 26,649). Sim et al. (2021) obtain their best performance on their Dupdiv dataset with the Cylindrical Minkowski + Triple Fermi-Dirac (TFD). We compare it on our larger dataset with Cylindrical Minkowski + equation 6. The results are reported in Table 4 and show a consistent gain of 2% mean Average Precision by using a proper time separation function.

We use the following hyperparameters for Dupdiv: $M = R_1^d$, $\theta_1 = 0.4$, $\theta_2 = 0.07$, exponent $m = 0.5$, learning rate $= 0.02$, number of epochs $= 500$. 
Table 4: Link prediction for directed graphs. Median average precision (AP) percentages across 20 random initializations on a held-out test set.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Manifold dimensionality $d$</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>DREAM 5: Saccharomyces cerevisiae</td>
<td>33.0 ± 2.7</td>
<td>34.2 ± 2.8</td>
<td>40.2 ± 3.3</td>
<td>44.5 ± 3.5</td>
<td>49.0 ± 2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>29.2 ± 2.5</td>
<td>37.9 ± 1.3</td>
<td>46.9 ± 2.6</td>
<td>48.8 ± 1.4</td>
<td>47.9 ± 1.2</td>
<td></td>
</tr>
<tr>
<td></td>
<td>34.7 ± 2.2</td>
<td>38.6 ± 2.3</td>
<td>46.4 ± 3.1</td>
<td>52.7 ± 3.0</td>
<td>54.0 ± 2.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.2 ± 3.2</td>
<td>41.3 ± 2.1</td>
<td>49.9 ± 2.5</td>
<td>47.5 ± 3.1</td>
<td>49.4 ± 3.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>37.4 ± 3.2</td>
<td>42.7 ± 2.3</td>
<td>46.8 ± 3.5</td>
<td>53.4 ± 2.2</td>
<td>54.6 ± 2.1</td>
<td></td>
</tr>
<tr>
<td>Minkowski + equation 6</td>
<td>47.6 ± 1.1</td>
<td>51.3 ± 1.5</td>
<td>54.4 ± 1.1</td>
<td>54.7 ± 2.0</td>
<td>54.8 ± 1.3</td>
<td></td>
</tr>
<tr>
<td>Cylindrical Minkowski + equation 6</td>
<td><strong>50.0 ± 1.7</strong></td>
<td><strong>52.5 ± 1.4</strong></td>
<td><strong>55.2 ± 1.5</strong></td>
<td><strong>56.2 ± 1.4</strong></td>
<td><strong>55.7 ± 1.7</strong></td>
<td></td>
</tr>
<tr>
<td>de Sitter + equation 6</td>
<td>44.8 ± 2.1</td>
<td>51.6 ± 1.6</td>
<td><strong>55.6 ± 1.3</strong></td>
<td>55.3 ± 1.4</td>
<td>55.4 ± 1.4</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Manifold dimensionality $d$</th>
<th>3</th>
<th>5</th>
<th>10</th>
<th>50</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>DREAM 5: in silico</td>
<td>29.4 ± 2.1</td>
<td>32.9 ± 2.5</td>
<td>39.7 ± 1.8</td>
<td>39.8 ± 1.6</td>
<td>34.8 ± 1.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>28.8 ± 5.5</td>
<td>46.8 ± 4.6</td>
<td>50.8 ± 7.4</td>
<td>50.9 ± 1.5</td>
<td>52.5 ± 1.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>36.3 ± 2.3</td>
<td>43.1 ± 3.1</td>
<td>51.2 ± 3.0</td>
<td>57.7 ± 2.8</td>
<td>58.0 ± 2.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>38.1 ± 4.8</td>
<td>45.2 ± 2.3</td>
<td>51.9 ± 5.2</td>
<td>55.6 ± 4.2</td>
<td>56.0 ± 3.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>41.0 ± 3.6</td>
<td>48.4 ± 7.3</td>
<td>56.3 ± 8.4</td>
<td>58.9 ± 2.9</td>
<td>61.0 ± 1.9</td>
<td></td>
</tr>
<tr>
<td>Minkowski + equation 6</td>
<td>48.4 ± 1.2</td>
<td>49.4 ± 1.1</td>
<td>51.6 ± 1.2</td>
<td>58.1 ± 2.1</td>
<td>58.8 ± 1.1</td>
<td></td>
</tr>
<tr>
<td>Cylindrical Minkowski + equation 6</td>
<td><strong>52.5 ± 1.9</strong></td>
<td>56.5 ± 1.6</td>
<td>59.8 ± 1.5</td>
<td>60.4 ± 1.5</td>
<td>60.8 ± 1.3</td>
<td></td>
</tr>
<tr>
<td>de Sitter + equation 6</td>
<td>48.5 ± 1.9</td>
<td><strong>57.4 ± 1.5</strong></td>
<td><strong>62.0 ± 1.4</strong></td>
<td><strong>60.6 ± 1.6</strong></td>
<td><strong>61.1 ± 1.4</strong></td>
<td></td>
</tr>
</tbody>
</table>

D.3 Hierarchy Extraction

For both equation 7 and equation 8, we set $\theta = 10^{-2}$, $r = 1$ and train the model for $10^5$ iterations/epochs. In equation 8, the regularization parameter $\lambda$ is set to $\lambda = \frac{|E|}{|E^c|}$ where $|E|$ is the number of pairs that satisfy $(v_i, v_j) \in E$ and $|E^c|$ is the number of pairs that satisfy $(v_a, v_b) \notin E$.

We report scores for the NIPS dataset in Table 3. We report the Spearman rank correlation coefficient $\rho$ (Spearman, 1904) for all the authors (left), the authors with at least 10 coauthors (middle) and authors with at least 20 coauthors (right). Spacetimes return better performance for the subset of authors with at least 10 coauthors.
Table 5: Evaluation scores for the different learned representations (mean ± standard deviation). ↓ the lower the metric, the better. ↑ the larger the metric, the better.

<table>
<thead>
<tr>
<th>Manifold</th>
<th>Problem</th>
<th>(d(x, y))</th>
<th>Whole dataset (\rho (\uparrow))</th>
<th>Top (s_i \geq 10) (\rho (\uparrow))</th>
<th>Top (s_i \geq 20) (\rho (\uparrow))</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathbb{R}^4)</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.469</td>
<td>0.512</td>
<td>0.217</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.460</td>
<td>0.490</td>
<td>0.292</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.629</td>
<td>0.552</td>
<td>0.316</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.667</td>
<td>0.493</td>
<td>0.307</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.625</td>
<td>0.441</td>
<td>0.227</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 7</td>
<td>(d_1(x, y))</td>
<td>0.437</td>
<td>0.493</td>
<td>0.387</td>
</tr>
<tr>
<td>(\mathbb{S}^1(r))</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.369</td>
<td>0.536</td>
<td>0.663</td>
</tr>
<tr>
<td>(\mathbb{R}^3)</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.524</td>
<td>0.668</td>
<td>0.484</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.538</td>
<td>0.326</td>
<td>0.143</td>
</tr>
<tr>
<td>(\mathbb{S}^1(r))</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.373</td>
<td>0.498</td>
<td>0.618</td>
</tr>
<tr>
<td>(\mathbb{R}^4)</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.478</td>
<td>0.678</td>
<td>0.543</td>
</tr>
<tr>
<td>(\mathbb{P}^1(r))</td>
<td>equation 8</td>
<td>(\chi^2_1(x, y))</td>
<td>0.576</td>
<td>0.455</td>
<td>0.219</td>
</tr>
</tbody>
</table>

Figure 4: Coordinates of embeddings \([-x, x] \in \mathbb{P}^1(1) \subset \mathbb{R}^3\) learned with equation 8.
Figure 5: (top) Coordinates of 2-dimensional embeddings $x = (x_0, x_1)^\top$ learned with equation 8 when $\mathcal{M} = \mathbb{R}^2_1$. (bottom) Coordinates of the first three coordinates of embeddings $x = (x_0, x_1, x_2, x_3)^\top$ learned with equation 8 when $\mathcal{M} = \mathbb{S}^3_1(r)$. In Lorentz geometry, a timelike geodesic joining two points is the longest timelike curve in a given convex normal neighborhood. This translates in the high-level nodes $v_1$ and $v_{34}$ being the furthest from the rest of the nodes. The ground truth edges are plotted in yellow and the node color corresponds to the joined faction. A small number of spacelike edges are visible (those edges more than 45 degrees from vertical).
E EXTENDED RELATED WORK

To help explain the contributions of our work relative to prior art, we use this section to provide a more detailed comparison of our contributions to that of Sim et al. (2021). Sim et al. (2021) extended Clough & Evans (2017) to the anti-de Sitter space and Lorentz cylinder. Although our motivation is similar, our contributions are methodological, rely on a simpler use of the intuitions of general relativity and Lorentzian pre-length spaces, and provide an easier interpretation of the learned representations as we explain below. First, Sim et al. (2021) do not address clearly the case when there is no geodesic between pairs of points, and their optimization framework leads to a distance loss term with a zero gradient in this case, which make it difficult to optimize. Moreover, in Sim et al. (2021), the prediction of an arc between a pair of nodes is determined via a Triple Fermi-Dirac (TFD) probability function that accounts for the squared (Lorentzian) distance between the nodes, the time coordinate difference $\Delta t$ and its opposite value $-\Delta t$. In other words, TFD accounts for both the chronological future and past (with different weights) of a given node. The major methodological difference with Sim et al. (2021) is that we restrict the representation of nodes connected by an arc to belong to $I^+(x, V_x)$ where $V_x$ is a convex normal neighborhood. Although subtle, this difference makes the optimization and interpretation of results easier.

In general, the Lorentzian distance function from $x$ to $y$ on $\mathcal{M}$ is defined to be infinite (i.e., $\chi_\mathcal{M}(x, y) = +\infty$) if $\mathcal{M}$ is non-chronological and there exists a closed timelike curve joining $x$ and $y$. Moreover, the Hopf-Rinow theorem does not hold for spacetimes. We propose to work with convex normal neighborhoods, which allows us to restrict the existence of arcs between nodes to the existence of geodesics joining events. Our distance function $\chi_V$ is called a local distance function (see Definition 4.25 of Beem et al. (1996)) when its domain is restricted to a convex normal neighborhood. Moreover, $\chi_V$ is continuous and differentiable on $V_x \times I^+(x, V_x)$ (see Lemma 4.26 of Beem et al. (1996)). In some cases, it might be easier to optimize its squared function $\chi_V^2$ which is of class $C^\infty$ on $V_x \times V_x$ (see Theorem 2.6 of Minguzzi (2019)). It is worth noting that $V_x$ can be defined to be globally hyperbolic for any spacetime (see Theorem 2.7 of Minguzzi (2019)). This means that $V_x$ admits a Cauchy time function that can be used to define a time separation function $\tau$ (see explanation in Section 3.2), whose sign defines the direction of edges. Our framework shares similarities with Sim et al. (2021) when $\mathcal{M} = \mathbb{R}^4 = V_x$ because $\mathbb{R}^4$ is globally hyperbolic and any pair of points of $\mathbb{R}^4$ can be joined by a geodesic. However, the way we define the time separation $\tau$ (instead of using the same $\Delta t$) is different when $\mathcal{M} = L^4_t([C])$ because we restrict it to be calculated on the maximal convex normal neighborhood. We construct $\tau$ so that it is positive if $x_j \in I^+(x, V_x)$ and negative if $x_j \in I^-(x, V_x)$. We then enforce $\tau$ to be positive if we want an arc from $v_i$ to $v_j$. The fact that we work only with an open convex set instead of the whole manifold $\mathcal{M}$ is crucial because $x_j$ can belong to both the chronological future $I^+(x_j, \mathcal{M}) := \{ y \in \mathcal{M} : x_j \ll y \}$ and past $I^-(x_j, \mathcal{M}) := \{ y \in \mathcal{M} : y \ll x_j \}$ if the $\mathcal{M}$ is non-chronological. Using the entire manifold requires that the sign of $\Delta t$ is not as informative, as in Sim et al. (2021). Our approach allows us to determine the direction of the arc joining $v_i$ and $v_j$ only via the sign of $\tau$.

We also define a general way of optimizing $\tau$ via the parallel transport (see Appendix 3.2) instead of working only with Cartesian coordinates, which is not meaningful for some spacetimes such as $\mathbb{R}^4_t(r)$. Moreover, since we restrict our Lorentzian distances to be calculated in the convex normal neighborhood $V_x$, we also have the nice interpretation that the Lorentzian distance corresponds to the length of the longest causal curve joining points.

One other contribution is the connection of our work with the theory of Lorentzian pre-length spaces (Kunzinger & Sämann, 2018) which does not require notions of differential geometry to be understood and can be applied to discrete topological spaces (see Example 2.16 of Kunzinger & Sämann (2018)). The framework proposed by Sim et al. (2021) is not a Lorentzian pre-length space due to their formulation of their time coordinate difference function that does not satisfy the properties of a time separation function (especially when $\mathcal{M}$ is non-chronological).
Algo 1 Pseudo-Riemannian optimization
input: differentiable function $f : M \to \mathbb{R}$ to be minimized, some initial value of $x \in M$

1: while not converge do
2: Calculate $\nabla f(x)$ \hfill i.e., $\nabla f(x)$ is the Euclidean gradient of $f$ at $x$ in the Euclidean ambient space
3: $\chi \leftarrow \Pi_x (G_1 \nabla f(x))$
4: $x \leftarrow \exp_x (-\eta \chi)$ \hfill where $\eta > 0$ is a step size
5: end while

F OPTIMIZATION

The optimizers that we use in the main paper for the different spacetimes can all be optimized as described in Algorithm 1. The goal is to minimize some function $f : M \to \mathbb{R}$ by using differential geometry tools described in [Gao et al., 2018; Law, 2021; Law & Stam, 2020]. $\nabla f(x)$ is the Euclidean gradient of $f$ at $x$, $\Pi_x (z)$ is the orthogonal projection of an arbitrary vector $z$ onto $T_x M$, and $G$ is an involutory matrix (i.e., $G^{-1} = G$). We consider in the following that the step size $\eta > 0$ is fixed and given. We give details for the different spacetimes that we consider in the main paper.

F.1 Minkowski spacetime $\mathbb{R}^d_1$

This optimizer was explained in [Gao et al., 2018]. When $M = \mathbb{R}^d_1$, $\Pi_x$ is the identity function. $G$ is the diagonal matrix with the first diagonal element equal to $-1$ and the remaining ones equal to $1$. $\exp_x(y) := x + y$. Algorithm 1 corresponds to the standard Euclidean gradient descent because $\chi = \nabla f(x)$ in this case.

F.2 Cylindrical Minkowski spacetime $\mathbb{L}^d_1(C)$

We recall that $\mathbb{L}^d_1(C) = \mathbb{R}^d_1 / \sim$, a quotient set defined such that $x \in \mathbb{R}^d_1$ and $y \in \mathbb{R}^d_1$ are equivalent (i.e., $x \sim y$) iff $\forall i > 0, y_i = x_i$, and $\exists k \in \mathbb{Z}, y_0 = x_0 + kC$ where $C > 0$ is a circumference hyperparameter.

When $M = \mathbb{L}^d_1(C)$, we use the same optimizer as in Section F.1.

Although it is optional, we also reproject the time coordinate of $x = (x_0, \ldots, x_{d-1})^\top$ at the end of each iteration as follows: $x_0 \leftarrow \left( ((x_0 + \frac{C}{2}) \mod C) - \frac{C}{2} \right) \in [-\frac{C}{2}, \frac{C}{2})$ where we use the modulo operation for real values which can be written as follows: $a \mod b := a - b \cdot \lfloor \frac{a}{b} \rfloor$, and $\lfloor \cdot \rfloor$ is the floor function. If the initial value of $x_0$ is not in $[-\frac{C}{2}, \frac{C}{2})$, this projects $x$ to an equivalent point.

F.3 De Sitter space $\mathbb{S}^d_1(r)$

This optimizer was introduced in [Law & Stam, 2020]. We recall that $\mathbb{S}^d_1(r) := \{ x \in \mathbb{R}^{d+1} : \langle x, x \rangle_1 = r^2 \}$. $G$ is the diagonal matrix with the first diagonal element equal to $-1$ and the remaining ones equal to $1$. We have: $\Pi_x(z) := z - \frac{\langle z, x \rangle_1}{\langle x, x \rangle_1} x = z - \frac{\langle z, x \rangle_1}{r^2} x$. The exponential map is defined in equation (10).

F.4 Anti-de Sitter space $\mathbb{H}^d_1(r)$

We recall that $\mathbb{H}^d_1(r) := \{ x \in \mathbb{R}^{d+1} : \langle x, x \rangle_2 = -r^2 \}$. $G$ is the diagonal matrix with the first two diagonal elements equal to $-1$ and the remaining ones equal to $1$. We have: $\Pi_x(z) := z - \frac{\langle z, x \rangle_2}{\langle x, x \rangle_2} x = z + \frac{\langle z, x \rangle_2}{r^2} x$. The exponential map is defined in equation (15).

F.5 Projective version of the anti-de Sitter space $\mathbb{P}^d_1(r)$

The neural network optimizer is given in [Law, 2021]. Otherwise, the embedding optimizer is the same as in Section F.4. The main difference is how the points are compared to calculate the distance and time separation function.


G  EXAMPLES OF GRAPHS

Since our general framework is fairly abstract, we give some explicit examples of directed graphs with or without cycles that can be described by our framework.

G.1  DIRECTED ACYCLIC GRAPHS (DAGs)

![DAG containing an undirected cycle](image)

Figure 6: DAG containing an undirected cycle.

We first consider a directed acyclic graph that consists of an undirected cycle (Figure 6).

The DAG $G = (V, E)$ is defined as $V = \{v_i\}_{i=1}^{n}$ and the set of arcs is $E = \{(v_1, v_4), (v_2, v_4), (v_2, v_5), (v_3, v_5), (v_1, v_6), (v_3, v_6)\}$. One can see that $G$ is not a directed tree since the undirected path $v_1, v_4, v_2, v_5, v_3, v_6, v_1$ is cyclic.

We recall that if $M$ is globally hyperbolic, $M$ is also chronological and can then only describe DAGs with our framework. Let us give an example where $M = \mathbb{R}^{d+1}_1$ or $M = S^d_1(r) \subset \mathbb{R}^{d+1}_1$ with $d = 3$ and $r > 0$ (e.g., $r = 1$). Both $\mathbb{R}^{d+1}_1$ and $S^d_1(r)$ are globally hyperbolic. Let us consider any value $\varepsilon > 0$, $a = r + \varepsilon$ and $b = \sqrt{2a^2 - r^2}$. We take the embeddings: $x_1 = (0, r, 0)^T, x_2 = (0, 0, 0, r)^T, x_3 = (b, a, a, 0)^T, x_5 = (b, 0, a, a)^T, x_6 = (b, a, 0, a)^T$.

When $M = S^d_1(r)$, we know from Section B.2 that $\bar{x}_1, \bar{x}_2$ is timelike (i.e., $\langle \bar{x}_1, \bar{x}_2 \rangle < 0$) iff $(x_1, x_2)^T > r^2$. By using the time separation function in equation [33] and assuming $\bar{x}_1, \bar{x}_2$ is timelike, we only need to compare the first coordinate of $x_i$ and $x_j$ to determine the direction of the edge between $v_i$ and $v_j$.

The case $M = \mathbb{R}^{d+1}_1$ is similar except that $\bar{x}_1, \bar{x}_2 := x_j - x_i$ is timelike iff $(\bar{x}_1, \bar{x}_2)^T < 0$.

One can verify that $\forall i,j,(v_i, v_j) \in E \iff x_j \in U_k$ and the tangent vector $\bar{x}_i, \bar{x}_j$ is future-directed timelike.

G.2  GRAPHS WITH DIRECTED CYCLES

G.2.1  MINKOWSKI CYLINDER $L^d_1(C)$

![A simple directed cycle](image)

Figure 7: A simple directed cycle.

We now illustrate a simple example of graph with directed cycle that can be drawn with $L^d_1(C)$ (Figure 2). The graph $G = (V, E)$ is defined as $V = \{v_i\}_{i=1}^{4}$ and $E = \{(v_1, v_2), (v_2, v_3), (v_3, v_1)\}$, which is a graph with directed cycle.

Let us consider that $C = 3$ and $d = 2$. The maximal normal neighborhood of every point $x \in \mathbb{R}^d$ is $U_k = \{y = (y_0, \ldots, y_{d-1})^T \in \mathbb{R}^d : y_0 \in (x_0 - 1.5, x_0 + 1.5)\}$. We also consider that $V_x = U_k$.

Since $L^d_1(C)$ is a quotient set, its points are equivalence classes. We consider three equivalence classes $[x_i] := \{(i + 3k, 0)^T : k \in \mathbb{Z}\}$ where $i \in \{1, 2, 3\}$ and we define $x_i := (i, 0)^T \in \mathbb{R}^d$. In other words, the five points $x_0 = (0, 0)^T, x_1 = (1, 0)^T, x_2 = (2, 0)^T, x_3 = (3, 0)^T, x_4 = (4, 0)^T$ in $\mathbb{R}^d$ actually correspond to three points in $L^d_1(C)$ due to the equivalence relation (i.e., $x_3 \sim x_0$ and $x_4 \sim x_1$). We can then compare those three equivalence classes by comparing $x_i$ with $x_{i-1}$ and $x_{i+1}$ only. One can verify that $\forall i \in \{1, 2, 3\}$, we have $x_{i+1} \in U_{x_i}, x_{i+1} \in I^+(x_i, U_{x_i}),$ and $x_{i-1} \in U_{x_i}$. However, we also have $x_{i-1} \notin I^+(x_i, U_{x_i})$. We then obtain the graph illustrated in Figure 4.
G.2.2 Projective version of the anti-de Sitter space $\mathbb{P}_d^d(r)$

We now consider the graph $G = (V, E)$ defined as $V = \{v_i\}_{i=1}^4$ and $E = \{(v_1, v_4), (v_2, v_1), (v_3, v_1), (v_3, v_2), (v_4, v_2), (v_4, v_3)\}$. This is a graph with directed cycles (e.g., $v_1 \rightarrow v_4 \rightarrow v_2 \rightarrow v_1$, see Figure 8).

We recall that every point of $\mathbb{P}_d^d(r)$ can be written as the unordered pair $\{-x, x\}$ where $x \in \mathbb{H}_d^d(r) \subset \mathbb{R}_2^{d+1}$. Taking $d = 2$, the maximal normal neighborhood of every point $\{-x, x\}$ can be written $U_{\{-x, x\}} = \{\{-y, y\} : y \in \mathbb{H}_d^d(r), \langle x, y \rangle_2 < 0\}$.

Let us consider $r = 1, \varepsilon_1 = -0.1, \varepsilon_2 = \varepsilon_3 = -0.5, a = -1, b = -\sqrt{(r + \varepsilon_1)^2 - r^2 + a^2}, c = -2, e = -\sqrt{(r + \varepsilon_2)^2 - r^2 + c^2}, g = 1, h = \sqrt{(r + \varepsilon_3)^2 - r^2 + g^2}$. We construct the four following points: $x_1 = p = (r, 0, 0)^{\top}, x_2 = (r + \varepsilon_1, a, b)^{\top}, x_3 = (r + \varepsilon_2, c, e)^{\top}, x_4 = (r + \varepsilon_3, g, h)^{\top}$. To define the future direction, we consider the timelike tangent vector $t = (0, 1, 0)^{\top} \in T_p\mathbb{H}_d^d(r)$.

In our framework, there exists an edge between $x_i$ and $x_j$ iff $|\langle x_i, x_j \rangle_2| \in (0, r^2)$. The direction of the edge is determined by using Section C.2.4.

Figure 8: A graph with directed cycles.