Generalization Error Bound for Hyperbolic Ordinal Embedding

Atsushi Suzuki [1] **Atsushi Nitanda** [2] **Jing Wang** [1] **Linchuan Xu** [3] **Marc Cavazza** [1] **Kenji Yamanishi** [4]

Abstract

Hyperbolic ordinal embedding (HOE) represents entities as points in hyperbolic space so that they agree as well as possible with given constraints in the form of entity i is more similar to entity j than to entity k. It has been experimentally shown that HOE can obtain representations of hierarchical data such as a knowledge base and a citation network effectively, owing to hyperbolic space's exponential growth property. However, its theoretical analysis has been limited to ideal noiseless settings, and its generalization error in compensation for hyperbolic space's exponential representation ability has not been guaranteed. The difficulty is that existing generalization error bound derivations for ordinal embedding based on the Gramian matrix are not applicable in HOE, since hyperbolic space is not inner-product space. In this paper, through our novel characterization of HOE with decomposed Lorentz Gramian matrices, we provide a generalization error bound of HOE for the first time, which is at most exponential with respect to the embedding space's radius. Our comparison between the bounds of HOE and Euclidean ordinal embedding shows that HOE's generalization error comes at a reasonable cost considering its exponential representation ability.

1. Introduction

Ordinal embedding, also known as a non-metric multidimensional scaling (Shepard, 1962a;b; Kruskal, 1964a;b), aims to represent entities as points in a metric space so that they are as consistent as possible with given ordinal data in the form of "entity i is more similar to entity j than to entity k." Many ordinal embedding methods have been proposed to obtain representations in Euclidean space, which we call Euclidean ordinal embedding (EOE) in this paper, and their effectiveness has been shown in a variety of machine learning areas such as embedding image data and artist dataset (Agarwal et al., 2007; Tamuz et al., 2011; van der Maaten & Weinberger, 2012). However, Euclidean space has a limitation in embedding data with a hierarchical tree-like structure (Lamping & Rao, 1994; Ritter, 1999; Nickel & Kiela, 2017) such as knowledge bases and complex networks. This limitation is due to Euclidean space's polynomial growth property, which means that the volume or surface of a ball in Euclidean space grows polynomially with respect to its radius. This Euclidean space's growth speed is significantly slower than what embedding hierarchical data such as an r-ary tree ($r \geq 2$) requires, which is exponential. To overcome this limitation, a few recent papers (Suzuki et al., 2019; Tabaghi & Dokmanic, 2020) have proposed ordinal embedding methods using hyperbolic space for hierarchical data, which we call hyperbolic ordinal embedding (HOE) in this paper. In contrast to Euclidean space's polynomial growth property, hyperbolic space has the exponential growth property, that is, the volume of any ball in hyperbolic space grows exponentially with respect to its radius (Lamping & Rao, 1994; Ritter, 1999; Nickel & Kiela, 2017). As a result, we can embed any tree to hyperbolic space with arbitrarily low distortion (Sarkar, 2011), and conversely, hyperbolic space's distance structure is well-approximated by a tree. By leveraging this hyperbolic space's advantage, Suzuki et al. (2019) have proposed an HOE method based on Riemannian stochastic gradient descent and achieved effective embedding of hierarchical tree-like data in low-dimensional space. Recently, Tabaghi & Dokmanic (2020) have solved the hyperbolic distance geometry problem, which includes HOE as a special case, by semi-definite relaxation of the problem and projection operation from Minkowski space to a hyperboloid. These two papers have experimentally shown HOE's potential ability to obtain low-dimensional representations effectively for hierarchical tree-like data such as a knowledge base and a citation network. However, the theoretical guarantee of HOE's performance is limited to ideal noiseless settings (Suzuki et al., 2019), and HOE's generalization performance in general noisy settings has not

[1]School of Computing and Mathematical Sciences, Faculty of Liberal Arts and Sciences, University of Greenwich, United Kingdom. [2]Department of Artificial Intelligence, Faculty of Computer Science and Systems Engineering, Kyushu Institute of Technology, Japan. [3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong. [4]Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo, Japan. Correspondence to: Atsushi Suzuki <atsushi.suzuki.rd@gmail.com>, Jing Wang <jing.wang@greenwich.ac.uk>.

Proceedings of the 38th International Conference on Machine Learning, PMLR 139, 2021.

been theoretically guaranteed, although HOE could have much worse generalization error than EOE in compensation for hyperbolic space's exponential growth property and cause overfitting for real data, which are often noisy.

In this paper, we derive the generalization error bound of HOE in general noisy settings under direct conditions on the radius of the embedding space. To the best of our knowledge, this is the first work that derives a generalization error bound for HOE. Whereas the generalization error of a learning model reflects the volume of its hypothesis space, owing to hyperbolic space's exponential growth property, we cannot expect HOE to have linear or polynomial generalization error with respect to the embedding space's radius, although it is proved for EOE by Jain et al. (2016) reflecting Euclidean space's polynomial growth property. Hence, our objective is to clarify the dependency of the error bound on the embedding space's radius as well as the number of entities and the size of ordinal data. In this paper, we show that HOE's generalization error is at most exponential with respect to the embedding space's radius. Also, the bound's dependency on the number of entities and the size of ordinal data is the same up to constant factors as that of EOE. Comparing our bound and that of EOE, we see that we can formally obtain HOE's bound by replacing a linear term in EOE's bound with respect to the embedding space's radius by an exponential term. This means that the generalization error bounds of HOE and EOE reflect the volume of their embedding space, and our HOE bound is reasonable as a cost for HOE's exponential representation ability.

The difficulty of deriving HOE's generalization error bound is that the technique for EOE's bound of formulating EOE's model and the restriction on its embedding space by the Gramian matrix (Jain et al., 2016) does not work for HOE. This is because hyperbolic space is not inner-product space and the Gramian matrix does not reflect its metric structure on which the HOE model is constructed. We solve this problem by our novel characterization of HOE model and the restriction on its embedding space by the decomposed Lorentz Gramian matrices. By our approach, we can formulate HOE model as a linear prediction model where the decomposed Lorentz Gramian matrices work as parameters and the restriction on its embedding space as conditions on the norms of these matrices. The resulting formulation enables us to calculate the Rademacher complexity (Koltchinskii, 2001; Koltchinskii & Panchenko, 2000; Bartlett et al., 2002) of the HOE model, which gives us a tight generalization error bound for linear prediction models (Kakade et al., 2008). Combining our Rademacher complexity calculation with existing standard statistical learning theory method (Bartlett & Mendelson, 2002), we obtain our HOE's generalization error bound.

1.1. Our Contributions

We derive the generalization error bound for HOE for the first time. Our bound is valid under intuitive conditions on the embedding space's radius and the simple uniform distribution assumption on ordinal data. Our results show that HOE's generalization error is at most exponential with respect to embedding space's radius. The bound's dependency on the number of entities and the size of ordinal data is the same as that of EOE up to constant factors.

1.2. Related Work

By leveraging hyperbolic space's exponential growth property, many papers have proposed embedding models using hyperbolic space in a variety of areas such as interactive visualization (Lamping & Rao, 1994; Walter & Ritter, 2002), Internet graph representations (Shavitt & Tankel, 2008; Boguná et al., 2010), routing problems in geographic communication networks (Kleinberg, 2007), and modeling complex networks (Krioukov et al., 2010). Recently, hyperbolic space has also attracted attention in many areas of machine learning such as graph embedding (Nickel & Kiela, 2017; Ganea et al., 2018a), metric multi-dimensional scaling (Sala et al., 2018), neural networks (e.g., Ganea et al., 2018b; Chami et al., 2019; Gülçehre et al., 2019) (See also Peng et al., 2021), word embedding (Tifrea et al., 2019), and multi-relational graph embedding (Suzuki et al., 2018; Balazevic et al., 2019), whereas machine learning methods for data in hyperbolic space have also been proposed (Cho et al., 2019; Chami et al., 2020). Ordinal embedding using hyperbolic space has also been proposed recently (Suzuki et al., 2019; Tabaghi & Dokmanic, 2020), whereas ordinal embedding has been originally studied intensively in Euclidean settings (Agarwal et al., 2007; Tamuz et al., 2011; van der Maaten & Weinberger, 2012; Terada & von Luxburg, 2014; Hashimoto et al., 2015; Cucuringu & Woodworth, 2015; Ma et al., 2018; Anderton & Aslam, 2019; Ma et al., 2019). From a theoretical perspective, the low distortion property of hyperbolic space for embedding a tree has been discussed in (Sarkar, 2011; Sala et al., 2018; Suzuki et al., 2019) under noiseless conditions. However, to the best of our knowledge, the generalization error in noisy settings of a machine learning model using hyperbolic space as embedding space has not been analyzed.

Rademacher complexity (Koltchinskii, 2001; Koltchinskii & Panchenko, 2000; Bartlett et al., 2002) is one of the key tools to derive an upper bound for the generalization error of a learning model. The upper bound derivation using Rademacher complexity has been studied in e.g., (Koltchinskii & Panchenko, 2002; Bartlett & Mendelson, 2002). The Rademacher complexity of a linear prediction model under norm restrictions has intensively been studied in (Kakade et al., 2008). Recently, Jain et al. (2016) calculated an upper bound of the Rademacher complexity of an EOE model. However, the Rademacher complexity of an HOE model has not been evaluated, and this paper contains a first presentation of such an evaluation.

2. Preliminaries

Notation In this paper, the symbol $:=$ is used to state that its left hand side is defined by its right hand side. We denote by $\mathbb{Z}, \mathbb{Z}_{>0}, \mathbb{R}, \mathbb{R}_{\geq 0}$ the set of integers, the set of positive integers, the set of real numbers, and the set of non-negative real numbers, respectively. Suppose that $D, N \in \mathbb{Z}_{>0}$. We denote by , $\mathbb{R}^D, \mathbb{R}^{D,N}, \mathbb{S}^{N,N}$ the set of D-dimensional real vectors, the set of real matrices with the size of $D \times N$, and the set of $N \times N$ symmetric matrices, respectively. For $N \in \mathbb{Z}$, $[N]$ denotes the set $\{1, 2, \ldots, N\}$ of integers. For a matrix $\boldsymbol{A} \in \mathbb{R}^{D,N}$, we denote by $[\boldsymbol{A}]_{d,n}$ the element in the d-row and the n-th column and by $\mathrm{Tr}(\boldsymbol{A})$ the trace of $\boldsymbol{A}$. For a vector $\boldsymbol{x} \in \mathbb{R}^D$, we denote by $\|\boldsymbol{x}\|_2$ the ℓ^2-norm of $\boldsymbol{x}$, defined by $\|\boldsymbol{x}\|_2 = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$. For matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{D,N}$, we denote by $\langle \boldsymbol{A}, \boldsymbol{B} \rangle_{\mathrm{F}}$ the Frobenius inner-product of $\boldsymbol{A}$ and $\boldsymbol{B}$, defined by $\mathrm{Tr}\left(\boldsymbol{A}^\top \boldsymbol{B}\right)$. For symmetric matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{S}^{N,N}$, we write $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite.

2.1. Ordinal Embedding

First, we formulate the ordinal embedding, which is of interest in this paper. Let N be the number of entities and we identify the set $[N]$ with the N entities. We assume that there exists a true dissimilarity measure $\Delta^* : [N] \times [N] \to \mathbb{R}_{\geq 0}$, where $\Delta^*(i,j)$ indicates the true dissimilarity between entity i and entity j. Ordinal data is a set of ordinal comparisons in the form of entity i is more similar to entity j than to entity k, which indicates $\Delta^*(i,j) < \Delta^*(i,k)$ if there is no noise in the comparison. In the following, we formulate ordinal comparisons according to the formulation by Jain et al. (2016). The s-th comparison consists of a pair of a triplet $(i_s, j_s, k_s) \in \mathcal{T}$ and a label $y_s \in \{-1, +1\}$, where

$$\mathcal{T} := \{(i,j,k) | i,j,k \in [N], j < k, k \neq i \neq j\}. \tag{1}$$

In (1), we impose the constraint $j < k$ to keep the uniqueness of the formulation. Note that $|\mathcal{T}| = \frac{1}{2}N(N-1)(N-2)$. The label indicates the result of the ordinal comparison. Specifically, $y_s = -1$ indicates i_s is closer to j_s than to k_s, and $y_s = +1$ indicates the converse. If there is no noise in the comparison, $y_s = -1$ and $y_s = +1$ means $\Delta^*(i,j) < \Delta^*(i,k)$ and $\Delta^*(i,j) > \Delta^*(i,k)$, respectively. Note that we also consider noisy comparison cases in this paper. Also, we assume that $\Delta^*(i,j) \neq \Delta^*(i',j')$ holds for any two different pairs $(i,j), (i',j')$ of different entities, to avoid ambiguity in comparison, as implicitly assumed also in (Jain et al., 2016). Let $(\mathcal{Z}, \Delta_{\mathcal{Z}})$ be a metric space, where $\mathcal{Z}$ is a set $\Delta_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}_{\geq 0}$ is a distance function on $\mathcal{Z}$. The objective of ordinal embedding in $\mathcal{Z}$ is to get representations $z_1, z_2, \ldots, z_N$ in some low-dimensional metric space $\mathcal{Z}$, such that the representations are consistent to the true dissimilarity measure Δ^*, where $z_i \in \mathcal{Z}$ is the representation of entity $i \in [N]$. Specifically, ideal representations should satisfy the following:

$$\Delta^*(i,j) \lesseqgtr \Delta^*(i,k) \Leftrightarrow \Delta_{\mathcal{Z}}(z_i, z_j) \lesseqgtr \Delta_{\mathcal{Z}}(z_i, z_k), \tag{2}$$

for a new triplet $(i,j,k) \in \mathcal{T}$, which may be unseen in the training data. We call the metric space $(\mathcal{Z}, \Delta_{\mathcal{Z}})$ used in ordinal embedding the *embedding space*. As ordinal embedding represents the dissimilarity between two entities by the distance between the two representations, embedding space selection is essential. In the next section, we introduce hyperbolic space, which HOE use as the embedding space.

2.2. Hyperbolic Space

Hyperbolic space is a metric space, which has been widely used to represent hierarchical data in machine learning areas, owing to its exponential growth property. In this section, we give a formal definition of hyperbolic space. There exist many well-known models of hyperbolic space, such as the hyperboloid model, the Beltrami-Klein model, Poincaré ball model, and Poincaré half-space model (see, e.g., Lee, 2018, Chapter 3), which are isometrically isomorphic to each other. In this paper, we mainly work on the hyperboloid model, which formulates hyperbolic space as a submanifold of Minkowski space. The advantage of the hyperboloid model is that we can use the inner product function of Minkowski space for discussion, which plays key role in our novel characterization of HOE. See e.g., (Lee, 2018, Chapter 3) for details.

The $(1+D)$-dimensional Minkowski space $\mathbb{M}^{1,D} = \left(\mathbb{R}^D, \langle\cdot,\cdot\rangle_{\mathrm{M}}\right)$, where $\langle\cdot,\cdot\rangle_{\mathrm{M}} : \mathbb{M}^{1,D} \times \mathbb{M}^{1,D} \to \mathbb{R}$ is the Lorentz inner-product function defined by

$$\begin{aligned} &\left\langle \begin{bmatrix} x_0 & x_1 & \ldots & x_D \end{bmatrix}^\top, \begin{bmatrix} x'_0 & x'_1 & \ldots & x'_D \end{bmatrix}^\top \right\rangle_{\mathrm{M}} \\ &:= -x_0 x'_0 + \sum_{d=1}^{D} x_d x'_d, \end{aligned} \tag{3}$$

is the pseudo Euclidean vector space with signature $(1, D)$, that is, the $(1+D)$-dimensional vector space equipped with the non-positive-definite bilinear function $\langle\cdot,\cdot\rangle_{\mathrm{M}}$. The hyperboloid model of D-dimensional hyperbolic space is a Riemannian manifold $\left(\mathbb{L}^D, g_x\right)$ embedded in $(1+D)$-dimensional Minkowski space $\mathbb{M}^{1,D}$, where

$$\mathbb{L}^D := \left\{ \boldsymbol{x} \in \mathbb{R}^{1+D} \middle| \langle \boldsymbol{x}, \boldsymbol{x} \rangle_{\mathrm{M}} = -1, x_0 > 0. \right\}, \tag{4}$$

and g_x is induced by the inclusion $\iota : \mathbb{L}^D \to \mathbb{M}^{1,D} : \boldsymbol{x} \mapsto \boldsymbol{x}$. The distance function $\Delta_{\mathbb{L}^D} : \mathbb{L}^D \times \mathbb{L}^D \to \mathbb{R}_{\geq 0}$ is given by

$$\Delta_{\mathbb{L}^D}(\boldsymbol{x}, \boldsymbol{x}') := \operatorname{arcosh}(-\langle \boldsymbol{x}, \boldsymbol{x}' \rangle_{\mathrm{M}}), \tag{5}$$

where arcosh is the area hyperbolic cosine function, which is the inverse function of the hyperbolic cosine function. Hyperbolic space has the *exponential growth property* in that the volume and surface area of any ball in hyperbolic space exponentially grows with respect to its radius. For example, the circumference of any ball with a radius of R in two-dimensional hyperbolic space is given by $2\pi \sinh R = O(\exp R)$ in constrast to $2\pi R$ in two-dimensional Euclidean space. Owing to this property, in graph embedding setting, we can embed any tree with arbitrarily low distortion to hyperbolic space in graph embedding setting (Sarkar, 2011), and in ordinal embedding setting, we can get representations that satisfy all the ordinal constraints generated by any tree (Suzuki et al., 2019) if the ordinal data are noiseless.

2.3. HOE and EOE

In this section, we define HOE. We also formulate EOE for the later comparison between HOE and EOE. HOE and EOE are ordinal embedding using hyperbolic space and Euclidean space, respectively. Specifically, HOE is ordinal embedding to obtain representations in $\mathbb{L}^D$ that agree as well as possible with (2) with $\Delta_{\mathcal{Z}} = \Delta_{\mathbb{L}^D}$. Likewise, EOE is ordinal embedding to obtain representations in $\mathbb{R}^D$ that agree as well as possible with (2) with $\Delta_{\mathcal{Z}} = \Delta_{\mathbb{R}^D}$ where $\Delta_{\mathbb{R}^D}$ is defined by $\Delta_{\mathbb{R}^D}(\boldsymbol{x}, \boldsymbol{x}') \coloneqq \|\boldsymbol{x}' - \boldsymbol{x}\|_2$.

As an embedding scheme, we mainly focus on minimizing the *empirical risk function* defined below. Let $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ and $\psi : \mathbb{R} \to \mathbb{R}$ be increasing functions and call them the *loss function* and *dissimilarity transformation function*, respectively. The empirical risk function $\hat{\mathcal{R}}^{\mathrm{z}}_{\mathcal{S}} : (\mathcal{Z})^N \to \mathbb{R}_{\geq 0}$ is defined by

$$\hat{\mathcal{R}}^{\mathrm{z}}_{\mathcal{S}}\Big((\boldsymbol{z}_n)_{n=1}^N\Big) \coloneqq \frac{1}{S}\sum_{s=1}^{S} \phi\Big(-y_s h\Big(i_s, j_s, k_s; (\boldsymbol{z}_n)_{n=1}^N\Big)\Big), \tag{6}$$

where the hypothesis function $h\Big(\cdot; (\boldsymbol{z}_n)_{n=1}^N\Big) : \mathcal{T} \to \mathbb{R}$ is defined by

$$h\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^N\Big) \coloneqq \psi(\Delta(\boldsymbol{z}_i, \boldsymbol{z}_j)) - \psi(\Delta(\boldsymbol{z}_i, \boldsymbol{z}_k)), \tag{7}$$

for $(\boldsymbol{z}_n)_{n=1}^N \in (\mathcal{Z})^N$.

For HOE, if we set $\phi(x) = \max\{0, x+1\}$ and $\psi(x) = x$, the risk function (6) is reduced to that proposed by Suzuki et al. (2019). In the following discussion, we set $\psi(x) = \cosh(x)$ for HOE and $\psi(x) = x^2$ for EOE as in (Jain et al., 2016). Whereas we can extend the following discussion for the case where another function is used for ψ, the generalization error bound for that case is worse than that of $\cosh$ if we follow the discussion below.

We also define the *expected risk function*:

$$\mathcal{R}^{\mathrm{z}}\Big((\boldsymbol{z}_n)_{n=1}^N\Big) \coloneqq \mathbb{E}_{(i,j,k),y}\phi\Big(-yh\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^N\Big)\Big). \tag{8}$$

Fix $\mathcal{B} \subset (\mathcal{Z})^N$, and we define the empirical risk minimizer $(\hat{\boldsymbol{z}}_n)_{n=1}^N$ and expected risk minimizer $(\boldsymbol{z}^*_n)_{n=1}^N$ by

$$\begin{aligned}(\hat{\boldsymbol{z}}_n)_{n=1}^N &\coloneqq \operatorname*{argmin}_{(\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}} \hat{\mathcal{R}}^{\mathrm{z}}_{\mathcal{S}}\Big((\boldsymbol{z}_n)_{n=1}^N\Big), \\ (\boldsymbol{z}^*_n)_{n=1}^N &\coloneqq \operatorname*{argmin}_{(\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}} \mathcal{R}^{\mathrm{z}}\Big((\boldsymbol{z}_n)_{n=1}^N\Big).\end{aligned} \tag{9}$$

Our interest in this paper is the *excess risk* given by

$$\mathcal{R}^{\mathrm{z}}\Big((\hat{\boldsymbol{z}}_n)_{n=1}^N\Big) - \mathcal{R}^{\mathrm{z}}\Big((\boldsymbol{z}^*_n)_{n=1}^N\Big), \tag{10}$$

which shows the generalization error of ordinal embedding.

3. Finite Sample Generalization Bound for HOE

3.1. Assumptions on Data Generation

To discuss the generalization error, we need to determine the distribution of data generation. Similar to (Jain et al., 2016), we assume that training data $((i_s, j_s, k_s), y_s)$ are generated independently and identically according to the following distributions.

Assumption 1. We assume that the triplets are generated independently and identically, and the conditional distribution of the label y_s given the triplet is determined by the true dissimilarity between i_s and j_s, and that between i_s and j_s as follows:

$$\mathbb{P}[y_s = +1 | (i_s, j_s, k_s) = (i, j, k)] = f(\Delta^*(i, j) - \Delta^*(i, k)), \tag{11}$$

where $f : \mathbb{R} \to [0, 1]$ is a fixed function called the *link function* (Jain et al., 2016).

3.2. Restriction on Representation Domain

To derive a finite generalization bound, in general, it is necessary to restrict parameters (in embedding cases, representations) to a bounded domain (e.g., linear prediction models (Bartlett & Mendelson, 2002; Kakade et al., 2008), neural networks (Bartlett & Mendelson, 2002; Schmidt-Hieber et al., 2020)). In this section, we discuss our restriction on embedding space. For the derived generalization bound to be practical, the restriction should be simple and geometrically intuitive. We put the following simple restrictions on the embedding space with respect to its radius.

Definition 1. Let $\boldsymbol{z}_0 \coloneqq \begin{bmatrix}1 & 0 & \dots & 0\end{bmatrix} \in \mathbb{L}^D$. For

$R, \rho \in \mathbb{R}_{\geq 0}$, we define $\mathcal{B}_R, \mathcal{B}^\rho, \mathcal{B}^\rho_R, \subset \left(\mathbb{L}^D\right)^N$ by

$$\mathcal{B}_R \coloneqq \left\{ (\boldsymbol{z}_n)_{n=1}^N \middle| \forall n \in [N] : \Delta_{\mathbb{L}^D}(\boldsymbol{z}_0, \boldsymbol{z}_n) \leq R \right\},$$

$$\mathcal{B}^\rho \coloneqq \left\{ (\boldsymbol{z}_n)_{n=1}^N \middle| \frac{1}{N} \sum_{n=1}^N \cosh^2 \Delta_{\mathbb{L}^D}(\boldsymbol{z}_0, \boldsymbol{z}_n) \leq \cosh^2 \rho \right\}, \tag{12}$$

and $\mathcal{B}^\rho_R \coloneqq \mathcal{B}_R \cap \mathcal{B}^\rho$, respectively.

Note that, since $\mathcal{B}_R \subset \mathcal{B}^R$, $\mathcal{B}^R_R = \mathcal{B}^R$ is valid. We mainly use $\mathcal{B}_R$ as the embedding space restriction since it is the simplest and most practical, we also show later that we can achieve a lower generalization error bound if we can set a small ρ for the restriction given by $\mathcal{B}^\rho_R$.

3.3. Main Result: Generalization Error

In this section, we give our main results, the upper bound of HOE's Rademacher complexity, to obtain HOE's generalization error bound.

Theorem 1. *Assume that ϕ is L_ϕ-Lipschitz and bounded. Define*

$$c_\phi \coloneqq \sup \phi\left(h\left(t; (\boldsymbol{z}_n)_{n=1}^N\right), y\right) - \inf \phi\left(h\left(t; (\boldsymbol{z}_n)_{n=1}^N\right), y\right), \tag{13}$$

where $\sup$ *and* $\inf$ *are taken over all $t \in \mathcal{T}, y \in \{-1, +1\}, (\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}^\rho_R$. Let $(\hat{\boldsymbol{z}}_n)_{n=1}^N$ and $(\boldsymbol{z}^*_n)_{n=1}^N$ be the empirical and expected risk minimizer of HOE defined by (9) with $\mathcal{B} = \mathcal{B}^\rho_R$. Then with probability at least 1 - δ we have that*

$$\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}^*_n)_{n=1}^N\right) \leq 2L_\phi\left(\cosh^2 \rho + \sinh^2 \rho\right)\left(\sqrt{\frac{2N^2 \ln N}{S}} + \frac{N \ln N}{\sqrt{12}S}\right) + 2c_\phi \sqrt{\frac{2 \ln \frac{2}{\delta}}{S}}. \tag{14}$$

Moreover, if the triplet is generated uniformly, that is, for all $(i, j, k) \in \mathcal{T}$,

$$\mathbb{P}[(i_s, j_s, k_s) = (i, j, k)] = \frac{1}{|\mathcal{T}|}. \tag{15}$$

is valid, then with probability at least 1 - δ we have that

$$\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}^*_n)_{n=1}^N\right) \leq 2L_\phi\left(\cosh^2 \rho + \sinh^2 \rho\right)\left(\sqrt{\frac{2(N+1) \ln N}{S}} + \frac{N \ln N}{\sqrt{12}S}\right) + 2c_\phi \sqrt{\frac{2 \ln \frac{2}{\delta}}{S}}. \tag{16}$$

If we only know R and L_ϕ, we have the following.

Corollary 2. *Suppose that $(\hat{\boldsymbol{z}}_n)_{n=1}^N$ and $(\boldsymbol{z}^*_n)_{n=1}^N$ are the empirical and expected risk minimizer of HOE in $\mathcal{B}_R = \mathcal{B}^R_R$. Then with probability at least 1 - δ we have that*

$$\begin{aligned}&\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}^*_n)_{n=1}^N\right) \\ &\leq 2L_\phi\left(\cosh^2 R + \sinh^2 R\right)\left(\sqrt{\frac{2N^2 \ln N}{S}} + \frac{N \ln N}{\sqrt{12}S}\right) \\ &\quad + 4L_\phi \cosh^2 (2R) \sqrt{\frac{2 \ln \frac{2}{\delta}}{S}} \\ &= O\left(L_\phi (\exp R)^2 \left(\sqrt{\frac{N^2 \ln N}{S}} + \frac{N \ln N}{S} + \sqrt{\frac{\ln \frac{2}{\delta}}{S}}\right)\right).\end{aligned} \tag{17}$$

Moreover, if the triplet is generated uniformly, then with probability at least 1 - δ we have that

$$\begin{aligned}&\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}^*_n)_{n=1}^N\right) \\ &\leq 2L_\phi\left(\cosh^2 R + \sinh^2 R\right)\left(\sqrt{\frac{2(N+1) \ln N}{S}} + \frac{N \ln N}{\sqrt{12}S}\right) \\ &\quad + 4L_\phi \cosh^2 (2R) \sqrt{\frac{2 \ln \frac{2}{\delta}}{S}} \\ &= O\left(L_\phi (\exp R)^2 \left(\sqrt{\frac{N \ln N}{S}} + \frac{N \ln N}{S} + \sqrt{\frac{\ln \frac{2}{\delta}}{S}}\right)\right).\end{aligned} \tag{18}$$

Proof. We can set $\rho = R$ in Theorem 1 since $\mathcal{B}^R_R = \mathcal{B}^R$. All we need to determine is c_ϕ in Theorem 1. For all $\boldsymbol{z}_i, \boldsymbol{z}_j$, we have that $\Delta_{\mathbb{L}^D}(\boldsymbol{z}_0, \boldsymbol{z}_i) \leq R$ and $\Delta_{\mathbb{L}^D}(\boldsymbol{z}_0, \boldsymbol{z}_j) \leq R$. Since hyperbolic space is a metric space, we have that $\Delta_{\mathbb{L}^D}(\boldsymbol{z}_i, \boldsymbol{z}_j) \leq 2R$, from the triangle inequality. Hence $c_\phi \leq 2L_\phi \cosh^2 (2R)$, which completes the proof. □

Remark 1. Consider the case where the triplet distribution is given by (15). The dependency of HOE's generalization error bounds suggests that the ordinal data size S should be $O(N \ln N)$, which is much smaller than the number $|\mathcal{T}|$ of possible triplets $O\left(N^3\right)$ and even smaller than the number of entity pairs $O\left(N^2\right)$. These dependencies of HOE's bound are the same as those of EOE's bound. We discuss the dependency on R in the next section.

Before giving our proof of Theorem 1, we first discuss the difference between HOE and EOE in the next section.

3.4. Comparison between EOE and HOE

In this section, we compare our generalization error bound of HOE to EOE's bound derived by Jain et al. (2016). First, we introduce the Gramian matrix and its norms, as preliminaries to discuss the results by Jain et al. (2016). The Gramian matrix of the representations $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots, \boldsymbol{z}_N \in \mathbb{R}^D$ is defined by

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \dots & \boldsymbol{z}_N \end{bmatrix}^\top \begin{bmatrix} \boldsymbol{z}_1 & \boldsymbol{z}_2 & \dots & \boldsymbol{z}_N \end{bmatrix} \in \mathbb{R}^{N,N}. \tag{19}$$

We denote by $\|\cdot\|_*$ and $\|\cdot\|_{\max}$ the nuclear norm and max norm defined by

$$\|\boldsymbol{A}\|_* := \sum_{n=1}^{N} \sigma_n(\boldsymbol{A}), \quad \|\boldsymbol{A}\|_{\max} := \max_{i,j\in[N]} [\boldsymbol{A}]_{i,j}, \tag{20}$$

respectively, where $\sigma_n(\boldsymbol{G})$ is the n-th singular value of $\boldsymbol{G}$.

EOE's restriction on its embedding space is given by the following sets.

Definition 2. We define $\mathcal{E}_\gamma, \mathcal{E}^\lambda, \mathcal{E}_\gamma^\lambda, \subset \left(\mathbb{R}^D\right)^N$ as follows:

$$\begin{aligned} \mathcal{E}_\gamma &:= \left\{ (\boldsymbol{z}_n)_{n=1}^N \middle| \|\boldsymbol{G}\|_{\max} \le \gamma \right\}, \\ \mathcal{E}^\lambda &:= \left\{ (\boldsymbol{z}_n)_{n=1}^N \middle| \|\boldsymbol{G}\|_* \le \lambda \right\}, \\ \mathcal{E}_\gamma^\lambda &:= \mathcal{E}_\gamma \cap \mathcal{E}^\lambda. \end{aligned} \tag{21}$$

The generalization error bound for EOE is given as follows:

Theorem 3 (Theorem 1 in (Jain et al., 2016))**.** *Assume that ϕ is L-Lipschitz. Let $(\hat{\boldsymbol{z}}_n)_{n=1}^N$ and $(\boldsymbol{z}_n^*)_{n=1}^N$ be the empirical and expected risk minimizer of EOE defined by (9) with $\mathcal{B} = \mathcal{E}_\gamma^\lambda$. Then with probability at least 1 - δ we have that*

$$\begin{aligned} &\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right) \\ &\le 12\sqrt{2}L\frac{\lambda}{N}\left(\sqrt{\frac{N\ln N}{S}} + \frac{\sqrt{3}}{9}\frac{N\ln N}{S}\right) + 12\sqrt{2}L\gamma\sqrt{\frac{\ln\frac{2}{\delta}}{S}}. \end{aligned} \tag{22}$$

Theorem 3 is not easy enough to interpret as its restrictions on the embedding space given by Definition 2 is not intuitive. Our following lemma shows that these restrictions are nothing else but those of the radius of the embedding space.

Lemma 4. *Let $\boldsymbol{G}$ be the Gramian matrix of $\boldsymbol{z}_1, \boldsymbol{z}_2, \dots \boldsymbol{z}_N$. Then the followings are valid.*

$$\begin{aligned} \|\boldsymbol{G}\|_* &= \sum_{n=1}^{N} \left[\Delta_{\mathbb{R}^D}(\boldsymbol{0}, \boldsymbol{z}_n)\right]^2, \\ \|\boldsymbol{G}\|_{\max} &= \max_{n\in[N]} \left[\Delta_{\mathbb{R}^D}(\boldsymbol{0}, \boldsymbol{z}_n)\right]^2. \end{aligned} \tag{23}$$

In particular the following holds:

$$\|\boldsymbol{G}\|_* \le N \|\boldsymbol{G}\|_{\max}. \tag{24}$$

Proof. See Supplementary Materials. □

Applying Lemma 4, we have a simplified version of Theorem 3. We define $\mathcal{B}_R^{\mathbb{R}^D} \subset \left(\mathbb{R}^D\right)^N$ by

$$\mathcal{B}_R^{\mathbb{R}^D} := \left\{ (\boldsymbol{z}_n)_{n=1}^N \middle| \forall n \in [N] : \Delta_{\mathbb{R}^D}(\boldsymbol{0}, \boldsymbol{z}_n) \le R \right\}. \tag{25}$$

Corollary 5. *Suppose that $(\hat{\boldsymbol{z}}_n)_{n=1}^N$ and $(\boldsymbol{z}_n^*)_{n=1}^N$ are the empirical and expected risk minimizer in $\mathcal{B}_R^{\mathbb{R}^D}$, then with probability at least 1 - δ we have that*

$$\begin{aligned} &\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right) - \mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right) \\ &\le 12\sqrt{2}LR^2\left(\sqrt{\frac{N\ln N}{S}} + \frac{\sqrt{3}}{9}\frac{N\ln N}{S} + \sqrt{\frac{\ln\frac{2}{\delta}}{S}}\right) \\ &= O\left(L_\phi R^2\left(\sqrt{\frac{N\ln N}{S}} + \frac{N\ln N}{S} + \sqrt{\frac{\ln\frac{2}{\delta}}{S}}\right)\right) \end{aligned} \tag{26}$$

Remark 2. We can formally obtain our generalization error bound of HOE by replacing R^2 in (26) by $(\exp R)^2$ up to constant factor. This is consistent to the fact that the volume of a ball in Euclidean space and hyperbolic space is polynomial and exponential with respect to its radius. We can say that HOE pays cost of exponential generalization error in compensation for its exponential volume for embedding space.

In this section, we compare EOE and HOE's generalization errors, that is variances. In the next section, by combining the above discussion about the variance and existing discussion about bias, we clarify the condition on which HOE outperforms EOE.

3.5. Bias-variance trade-off: where does HOE outperform EOE?

As in other machine learning models, there is a bias-variance trade-off between HOE and EOE. The comparison between EOE and HOE bounds shows that HOE's generalization error bound is larger than that of EOE, suggesting that we should not use hyperbolic space if euclidean space can represent the dissimilarity among the entities successfully. However, it does not mean that HOE's expected loss is always larger than that EOE, because HOE may have lower minimum expected loss $\mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right)$ than EOE, as existing theoretical analyses have shown in noiseless settings (e.g., (Sarkar, 2011)). In fact, if the true dissimilarity measure Δ^* is given by the graph distance of a weighted tree, where euclidean space cannot represent their metric structure well, then HOE can perform better than EOE with sufficient number S of ordinal data as discussed in the following. Suppose that $[N]$ is the set of vertices of a weighted tree and $\Delta^*(i,j)$ is the graph distance between $i \in [N]$ and $j \in [N]$ on the weighted tree. We consider the ramp loss ϕ_{ramp} defined by

$$\phi_{\mathrm{ramp}}(x) \coloneqq \begin{cases} 0 & x \leq -1. \\ x+1 & -1 \leq x \leq 0. \\ 1 & x \geq 1. \end{cases} \tag{27}$$

Suppose that the link function is given by

$$f(x) = \begin{cases} \frac{1}{2} + \alpha & \text{if } x > 0, \\ \frac{1}{2} - \alpha & \text{if } x < 0, \end{cases} \tag{28}$$

where $\alpha \in \mathbb{R}_{>0}$. For any embedding model, the expected loss is equal to or larger than $\frac{1}{2} - \alpha$. For an embedding model to achieve the minimum loss $\frac{1}{2} - \alpha$, the number V of triplets violating (2) must be zero. Specifically, the expected loss is no smaller than $\left(\frac{1}{2} - \alpha\right) + 2\alpha \frac{V}{|\mathcal{T}|}$. Let $V_{\min}^{\mathbb{R}^2}$ and $V_{\min}^{\mathbb{L}^2}$ denote the minimum V attainable by EOE using $\mathbb{R}^2$ and HOE using $\mathbb{L}^2$, respectively. Assume that the true dissimilarity Δ^* is the graph distance of a weighted tree. Regarding HOE, we have that $V_{\min}^{\mathbb{L}^2} = 0$. This is achieved by $(1+\epsilon)$-distortion Delaunay embedding algorithm (Sarkar, 2011). We formally state this fact in the following.

Lemma 6. *Suppose that $[N]$ is the set of vertices of a weighted tree and $\Delta^* : [N] \times [N] \to \mathbb{R}_{\geq 0}$ is given by its graph distance, and assume that for i, j, i', j' such that $i \neq j$, $i' \neq j'$ and $\{i.j\} \neq \{i'.j'\}$, $\Delta^*(i,j) \neq \Delta^*(i',j')$ is valid. Then there exist representations $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots, \boldsymbol{z}_N \in \mathbb{L}^2$ that satisfy $\Delta_{\mathbb{L}^2}(\boldsymbol{z}_i, \boldsymbol{z}_j) - \Delta_{\mathbb{L}^2}(\boldsymbol{z}_{i'}, \boldsymbol{z}_{j'}) > 1$ for all $i, j, i', j' \in [N]$ such that $\Delta^*(i,j) > \Delta^*(i',j')$.*

Proof. See Supplementary Materials. □

Hence, if we take R sufficiently large, the minimum expected loss of HOE in $\mathbb{L}^2$ is $\frac{1}{2} - \alpha$. On the other hand, EOE cannot achieve this minimum value for some trees (e.g., all the trees that have a node with degree larger than or equal to 6). We discuss this in the following. In EOE, $\mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right) = \left(\frac{1}{2} - \alpha\right) + 2\alpha \frac{V_{\min}^{\mathbb{R}^2}}{|\mathcal{T}|}$. Here, $V_{\min}^{\mathbb{R}^2}$ is not smaller than the number of disjoint 6-star subgraphs in the graph, as stated in the following lemma.

Lemma 7. *Assume that the true dissimilarity Δ^* is the graph distance of a weighted tree. There exists at least one triplet that violates (2) for each 6-star subgraph in the original graph.*

Proof. See Supplementary Materials. □

In this case, as HOE's generalization error bound converges to zero as $S \to \infty$, the expected loss of HOE's empirical risk minimizer can perform better than all representations in EOE. Specifically, we have the following condition on which HOE outperforms EOE.

Proposition 8. *With probability $1-\delta$, HOE's $\mathcal{R}^{\mathrm{z}}\left((\hat{\boldsymbol{z}}_n)_{n=1}^N\right)$ is smaller than EOE's if*

$$S > O\left(\frac{(\exp R)^2 |\mathcal{T}|}{\alpha V_{\min}^{\mathbb{R}^2} N \ln N}\left(\sqrt{N \ln N} + \sqrt{\ln \frac{2}{\delta}}\right) + \frac{1}{\sqrt{N \ln N} + \sqrt{\ln \frac{2}{\delta}}}\right)^2. \tag{29}$$

3.6. Limitation

Whereas EOE bound by Jain et al. (2016) and our HOE clarify the dependency of EOE and HOE's generalization error on the embedding space's radius R, respectively, the dependency on the embedding space's dimension D could be improved. Both of these bounds do not directly depend on D. This is not consistent with our intuition that using low-dimensional space should give low generalization error. Jain et al. (2016) has discussed the dependency of EOE's bound on D by substituting $\lambda = \sqrt{D} N \gamma$, based on the following observation:

$$\|\boldsymbol{G}\|_* \leq \sqrt{D} \|\boldsymbol{G}\|_{\mathrm{F}} \leq \sqrt{D} N \|\boldsymbol{G}\|_\infty. \tag{30}$$

However, as we have seen in (24), we can directly prove $\|\boldsymbol{G}\|_* \leq N \|\boldsymbol{G}\|_\infty$. Hence there is no reason to consider condition $\lambda = \sqrt{D} N \gamma$. Deriving a tighter bound in terms of the dimension for general ordinal embedding could be future work.

4. Proof Techniques

In this section, we explain our techniques used to prove Theorem 1. Firstly, we explain why existing generalization error bound derivation for EOE does not work in HOE. Secondly, we introduce the first key idea to solve the problem, which is the reformulation of HOE's hypothesis function by the Lorentz Gramian matrix. Thirdly, we introduce the second key idea, which is the reformulation of the restriction on HOE's embedding space by our novel characterization of the decomposed Lorentz Gramian matrix. Lastly, we provide the sketch of the proof.

4.1. Difficulty in HOE and Our Solution

In this section, we explain the difficulty of deriving the generalization error bound and our solutions to prove Theorem 1. We first see the approach for the existing EOE case. Jain et al. (2016) succeeded in deriving the generalization error bound of EOE following the procedures below:

- Converting the hypothesis function to that of a linear prediction problem in the following form:

$$h\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^{N}\Big) = \langle \boldsymbol{G}, \boldsymbol{T}_{i,j,k} \rangle_{\mathrm{F}}\,, \tag{31}$$

where $\boldsymbol{T}_{i,j,k} \in \mathbb{S}^{N,N}$ is some matrix determined by i, j, k. Here the Gramian matrix works as the parameter of a linear prediction model.

- Calculating the Rademacher complexity (Koltchinskii, 2001; Koltchinskii & Panchenko, 2000; Bartlett et al., 2002, See Supplementary Materials for definition.) of EOE's hypothesis functions under the restriction on norms of the Gramian matrix given by Definition 2, which we have shown are equivalent to those on the radius of the embedding space (Lemma 4).

However, this procedure does not straightforwardly work for HOE, owing to following reasons. The first reason is that as hyperbolic space is not inner-product space, it is difficult to convert HOE's hypothesis function to that of a linear prediction model in the form of (31), as long as we use the Gramian matrix as its parameters. We solve this problem using the Lorentz Gramian matrix to reformulate HOE's hypothesis function, leveraging the fact that hyperbolic space is a sub-manifold of Minkowski space. The second reason is that, even though we can reformulate HOE's hypothesis function as that of a linear prediction model, that conversion also converts our simple restrictions given in Definition 1 to an unknown set of the Lorentz Gramian matrices. To solve this second problem, we give a clear characterization of the restrictions on embedding space's radius as conditions on the decomposed Lorentz Gramian matrices. Our characterization by the decomposed Lorentz Gramian matrix enables us to calculate the Rademacher complexity of HOE's hypothesis functions using similar techniques to that used in (Jain et al., 2016). Once we can calculate the Rademacher complexity, we can prove Theorem 1 by the standard statistical learning technique.

4.2. Reformulation of HOE Hypothesis Function by Lorentz Gramian Matrix

Recall that Minkowski space is equipped with the Lorentz inner-product function $\langle \cdot, \cdot \rangle_{\mathrm{M}}$. The Lorentz Gramian matrix $\boldsymbol{L}$ of representations $\boldsymbol{z}_0, \boldsymbol{z}_1, \dots, \boldsymbol{z}_N \in \mathbb{L}^D$ is defined by $[\boldsymbol{L}]_{i,j} = \langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle_{\mathrm{M}}$. If $\psi = \cosh$, we can reformulate HOE's hypothesis function using the Lorentz inner-product function as follows:

$$h\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^{N}\Big) = -\langle \boldsymbol{z}_i, \boldsymbol{z}_j \rangle_{\mathrm{M}} + \langle \boldsymbol{z}_i, \boldsymbol{z}_k \rangle_{\mathrm{M}}\,. \tag{32}$$

For $(i, j, k) \in \mathcal{T}$, we define $\boldsymbol{T}_{i,j,k} \in \mathbb{S}^{N,N}$ by

$$[\boldsymbol{T}_{i,j,k}]_{n,n'} = \begin{cases} -\frac{1}{2} & \text{if } (n, n') = (i, j), (j, i), \\ +\frac{1}{2} & \text{if } (n, n') = (i, k), (k, i), \\ 0 & \text{otherwise.} \end{cases} \tag{33}$$

Then we can convert $h\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^{N}\Big)$ as a linear function of the Lorentz Gramian as follows:

$$h\Big(i, j, k; (\boldsymbol{z}_n)_{n=1}^{N}\Big) = \langle \boldsymbol{L}, \boldsymbol{T}_{i,j,k} \rangle_{\mathrm{F}}\,. \tag{34}$$

Using (34), we redefine the empirical risk function as a function of the Lorentz Gramian as follows:

$$\mathcal{R}_{\mathcal{S}}^{\mathrm{G}}(\boldsymbol{L}) \coloneqq \frac{1}{S} \sum_{s=1}^{S} \phi\big(y_{i,j,k} \cdot \langle \boldsymbol{L}, \boldsymbol{T}_{i_s, j_s, k_s} \rangle_{\mathrm{F}}\big). \tag{35}$$

By definition, if $\boldsymbol{L}$ is the Lorentz Gramian matrix of $(\boldsymbol{z}_n)_{n=1}^{N}$, $\mathcal{R}_{\mathcal{S}}^{\mathrm{G}}(\boldsymbol{L}) = \hat{\mathcal{R}}_{\mathcal{S}}^{z}\Big((\boldsymbol{z}_n)_{n=1}^{N}\Big)$ is valid. According to the above reformulation of HOE risk using the Lorentz Gramian matrix, the risk bound is obtained if the range of values that the Lorentz Gramian matrix can take is specified. In the next section, we discuss the exact range.

4.3. Decomposition of the Lorentz Gramian Matrix

In Section 3.2, we have put the restrictions on the radius of HOE's embedding space, and in the previous section, we have reformulated HOE hypothesis function as the function of the Lorentz Gramian matrix. To calculate the Rademacher complexity of the set of the HOE hypothesis functions and derive a generalization error bound, it is necessary to characterize the original restrictions on embedding space given in Definition 1 by conditions on the Lorentz Gramian matrix. In this section, we give our novel characterization of the restrictions as conditions with respect to the decomposed Lorentz Gramian matrices. The following lemma gives conditions on the Lorentz Gramian matrix that is equivalent to the restrictions in Section 3.2.

Lemma 9. *Let $R \in \mathbb{R}_{\geq 0}$ and $\boldsymbol{L}, \boldsymbol{L}^-, \boldsymbol{L}^+ \in \mathbb{S}^{N,N}$. Define conditions* **(a)-(f)** *as follows:*

(a) $\begin{cases} \textbf{(a-)}\ \boldsymbol{L}^- \succeq \mathbf{O}, \\ \textbf{(a+)}\ \boldsymbol{L}^+ \succeq \mathbf{O}, \end{cases}$

(b) $\begin{cases} \textbf{(b-)}\ \operatorname{rank} \boldsymbol{L}^- = 1, \\ \textbf{(b+)}\ \operatorname{rank} \boldsymbol{L}^+ \leq D, \end{cases}$

(c) $\forall n \in [N] : \left[\boldsymbol{L}^+ - \boldsymbol{L}^-\right]_{n,n} = -1,$

(d) $\forall i, j \in [N] : \left[\boldsymbol{L}^+ - \boldsymbol{L}^-\right]_{i,j} \leq -1,$

(e) $\begin{cases} \textbf{(e-)}\ \left\|\boldsymbol{L}^-\right\|_{\max} \leq \cosh^2 R, \\ \textbf{(e+)}\ \left\|\boldsymbol{L}^+\right\|_{\max} \leq \sinh^2 R, \end{cases}$

(f) $\begin{cases} \textbf{(f-)}\ \left\|\boldsymbol{L}^-\right\|_{*} \leq N \cosh^2 \rho, \\ \textbf{(f+)}\ \left\|\boldsymbol{L}^+\right\|_{*} \leq N \sinh^2 \rho, \end{cases}$

where conditions **(a)**, **(b)**, **(e)**, *and* **(f)** *are conditions* **(a-)** *and* **(a+)**, **(b-)** *and* **(b+)**, **(e-)** *and* **(e+)**, *and* **(f-)** *and* **(f+)**, *respectively. Then*

(i) $\boldsymbol{L}$ *is the Lorentz Gramian matrix of a series of representations* $(\boldsymbol{z}_n)_{n=1}^N \in (\mathbb{L}^D)^N$ *if and only if there exist* $\boldsymbol{L}^-, \boldsymbol{L}^+ \in \mathbb{S}^{N,N}$ *such that* $\boldsymbol{L} = \boldsymbol{L}^+ - \boldsymbol{L}^-$ *and conditions* **(a)-(d)** *are satisfied.*

(ii) $\boldsymbol{L}$ *is the Lorentz Gramian matrix of a series of representations* $(\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}_R$ *if and only if there exist* $\boldsymbol{L}^-, \boldsymbol{L}^+ \in \mathbb{S}^{N,N}$ *such that* $\boldsymbol{L} = \boldsymbol{L}^+ - \boldsymbol{L}^-$ *and conditions* **(a)-(e)** *are satisfied.*

(iii) $\boldsymbol{L}$ *is the Lorentz Gramian matrix of a series of representations* $(\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}^\rho$ *if and only if there exist* $\boldsymbol{L}^-, \boldsymbol{L}^+ \in \mathbb{S}^{N,N}$ *such that* $\boldsymbol{L} = \boldsymbol{L}^+ - \boldsymbol{L}^-$ *and conditions* **(a)-(d),(f)** *are satisfied.*

(iv) $\boldsymbol{L}$ *is the Lorentz Gramian matrix of a series of representations* $(\boldsymbol{z}_n)_{n=1}^N \in \mathcal{B}_R^\rho$ *if and only if there exist* $\boldsymbol{L}^-, \boldsymbol{L}^+ \in \mathbb{S}^{N,N}$ *such that* $\boldsymbol{L} = \boldsymbol{L}^+ - \boldsymbol{L}^-$ *and conditions* **(a)-(f)** *are satisfied.*

Proof. See Supplementary Materials. □

We call the pair $\boldsymbol{L}^-$ and $\boldsymbol{L}^-$ the *decomposed Lorentz Gramian matrices*. Lemma 9 rephrases the geometric restrictions on hyperbolic space to conditions including those on the max norm and nuclear norm of the decomposed Lorentz Gramian matrices. The significance of Lemma 9 is that this rephrasing enables us to use techniques similar to those in (Jain et al., 2016), where they considered the restriction on those norms of the ordinary Gramian matrix. Note that the statement of Lemma 9 (i) is equivalent to that of Proposition 1 in (Tabaghi & Dokmanic, 2020) and we can regard (ii) and (iii) as extensions of (i). However, their proof of the necessity in (i) is incomplete, for which we give a complete proof. See the remark in Supplementary Materials for details.

4.4. Proof Sketch of Theorem 1

Lastly, we give a brief sketch of our proof of Theorem 1. By the decomposition of the Lorentz Gramian matrix, we have the following form of HOE's hypothesis function.

$$h\left(i, j, k; (\boldsymbol{z}_n)_{n=1}^N\right) = \left\langle \boldsymbol{L}^+ - \boldsymbol{L}^-, \boldsymbol{T}_{i,j,k} \right\rangle_{\mathrm{F}}, \tag{36}$$

with constraints on norms of $\boldsymbol{L}^+$ and $\boldsymbol{L}^-$ given by Lemma 9. This enables us to decompose the Rademacher complexity of HOE's hypothesis function into two terms. These decomposed terms can be evaluated in a similar way to that used in (Jain et al., 2016). See Section F in Supplementary Materials for the definition of Rademacher complexity, our upper bound of the Rademacher complexity of HOE's hypothesis functions and its proof, and the complete proof of Theorem 1.

Table 1. Triplet classification error (%).

Dataset	S	100	200	400	800	1600
R2	**EOE**	**40.45**	**38.62**	**38.22**	**33.06**	**30.86**
	HOE	45.50	42.58	39.23	33.50	32.01
Star	**EOE**	**40.02**	**39.70**	**38.87**	39.46	36.30
	HOE	43.61	43.17	41.04	**39.11**	**35.74**

5. Experiments

We compared EOE and HOE on two types of ordinal datasets (R2 and Star). The ordinal data are generated according to (15) and (11), where f is given by (28) with $\alpha = 0.25$ and Δ^* is defined by a metric. The metric for R2 is the distance matrix of 20 points in $\mathbb{R}^2$, where we expect both HOE and EOE to have small $\mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right)$. The metric for Star is the distance matrix of a 20-star graph with random weights, where we expect HOE to have smaller $\mathcal{R}^{\mathrm{z}}\left((\boldsymbol{z}_n^*)_{n=1}^N\right)$ than EOE since a star graph is a special tree. We set $\psi(x) = x^2$ for EOE and $\psi(x) = \cosh x$ for HOE. The ordinal data size is $S = 100, 200, 400, 800, 1600$. We have set both batch size and the number of epoch in stochastic gradient descent to 1000. The learning rate has been selected from $\{0.1, 1.0, 10.0\}$ by grid-search, following (Suzuki et al., 2019). We have run 10 times for each dataset and report the average error in Table 1. EOE obtains smaller errors than HOE in R2, owing to EOE's smaller excess risk. In Star, HOE shows larger errors than HOE for small Ss (100,200,400) but smaller errors than HOE for large Ss (800.1600). This is inline with the analysis in Section 3.4.

6. Conclusion

We have shown that HOE's generalization error is at most exponential with respect to the embedding space's radius R. Also, we have seen that the bound's dependency on the number of entities and the size of ordinal data is the same up to constant factors as that of EOE. Comparing our bound and that of EOE, we have seen that we can formally obtain HOE's bound by replacing a linear term in EOE's bound with respect to the embedding space's radius by an exponential term. The generalization error bounds of HOE and EOE reflect the volume of embedding space, and our HOE bound is reasonable as a cost for HOE's exponential representation ability. Our bias-variance trade-off discussion suggests that although we should not use hyperbolic space where Euclidean space can represent the true dissimilarity well, HOE's generalization performance cost is worth paying if the data has hierarchical tree structure, as we have seen through the tree example. Combined with existing analyses of hyperbolic embedding in noiseless settings, our generalization error analysis in general noisy settings provides a guide for embedding space selection in real applications.

References

Agarwal, S., Wills, J., Cayton, L., Lanckriet, G. R. G., Kriegman, D. J., and Belongie, S. J. Generalized non-metric multidimensional scaling. In *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics*, pp. 11–18, 2007.

Anderton, J. and Aslam, J. A. Scaling up ordinal embedding: A landmark approach. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 282–290, 2019.

Balazevic, I., Allen, C., and Hospedales, T. M. Multi-relational poincaré graph embeddings. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pp. 4465–4475, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Boucheron, S., and Lugosi, G. Model selection and error estimation. *Machine Learning*, 48(1-3): 85–113, 2002.

Boguná, M., Papadopoulos, F., and Krioukov, D. Sustaining the internet with hyperbolic mapping. *Nature communications*, 1(1):1–8, 2010.

Chami, I., Ying, Z., Ré, C., and Leskovec, J. Hyperbolic graph convolutional neural networks. In *Proceedings of the 33rd Conference on Neural Information Processing Systems*, pp. 4869–4880, 2019.

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. In *Proceedings of the 34th Conference on Neural Information Processing Systems*, 2020.

Cho, H., Demeo, B., Peng, J., and Berger, B. Large-margin classification in hyperbolic space. In *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1832–1840, 2019.

Cucuringu, M. and Woodworth, J. Ordinal embedding of unweighted knn graphs via synchronization. In *Proceedings of 25th IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2015.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic entailment cones for learning hierarchical embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 1632–1641, 2018a.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. In *Proceedings of the 32nd Conference on Neural Information Processing Systems*, pp. 5350–5360, 2018b.

Gülçehre, Ç., Denil, M., Malinowski, M., Razavi, A., Pascanu, R., Hermann, K. M., Battaglia, P. W., Bapst, V., Raposo, D., Santoro, A., and de Freitas, N. Hyperbolic attention networks. In *Proceedings of 7th International Conference on Learning Representations*, 2019.

Hashimoto, T. B., Sun, Y., and Jaakkola, T. S. Metric recovery from directed unweighted graphs. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, 2015.

Jain, L., Jamieson, K. G., and Nowak, R. D. Finite sample prediction and recovery bounds for ordinal embedding. In *Proceedings of the 30th Conference on Neural Information Processing Systems*, pp. 2703–2711, 2016.

Kakade, S. M., Sridharan, K., and Tewari, A. On the complexity of linear prediction: Risk bounds, margin bounds, and regularization. In *Proceedings of the 22nd Conference on Neural Information Processing Systems*, pp. 793–800, 2008.

Kleinberg, R. Geographic routing using hyperbolic space. In *Proceedings of the 26th IEEE International Conference on Computer Communications*, pp. 1902–1909, 2007.

Koltchinskii, V. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Koltchinskii, V. and Panchenko, D. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pp. 443–457. Springer, 2000.

Koltchinskii, V. and Panchenko, D. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of statistics*, 30(1):1–50, 2002.

Krioukov, D., Papadopoulos, F., Kitsak, M., Vahdat, A., and Boguná, M. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

Kruskal, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964a.

Kruskal, J. B. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115–129, 1964b.

Lamping, J. and Rao, R. Laying out and visualizing large trees using a hyperbolic space. In *Proceedings of the 7th ACM Symposium on User Interface Software and Technology*, pp. 13–14, 1994.

Lee, J. M. *Introduction to Riemannian manifolds*. Springer, 2018.

Ma, K., Zeng, J., Xiong, J., Xu, Q., Cao, X., Liu, W., and Yao, Y. Stochastic non-convex ordinal embedding with stabilized barzilai-borwein step size. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, pp. 3738–3745, 2018.

Ma, K., Xu, Q., and Cao, X. Robust ordinal embedding from contaminated relative comparisons. In *Proceedings of the 33rd AAAI Conference on Artificial Intelligence*, pp. 7908–7915, 2019.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. In *Proceedings of the 31st Conference on Neural Information Processing Systems*, pp. 6338–6347, 2017.

Peng, W., Varanka, T., Mostafa, A., Shi, H., and Zhao, G. Hyperbolic deep neural networks: A survey. *arXiv preprint arXiv:2101.04562*, 2021.

Ritter, H. Self-organizing maps on non-euclidean spaces. In *Kohonen maps*, pp. 97–109. Elsevier, 1999.

Sala, F., Sa, C. D., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 4457–4466, 2018.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *Proceedings of the 19th International Symposium on Graph Drawing*, pp. 355–366, 2011.

Schmidt-Hieber, J. et al. Nonparametric regression using deep neural networks with relu activation function. *Annals of Statistics*, 48(4):1875–1897, 2020.

Shavitt, Y. and Tankel, T. Hyperbolic embedding of internet graph for distance estimation and overlay construction. *IEEE/ACM Transactions on Networking*, 16(1):25–36, 2008.

Shepard, R. N. The analysis of proximities: multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27(2):125–140, 1962a.

Shepard, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika*, 27(3):219–246, 1962b.

Suzuki, A., Enokida, Y., and Yamanishi, K. Riemannian transe: Multi-relational graph embedding in non-euclidean space, 2018.

Suzuki, A., Wang, J., Tian, F., Nitanda, A., and Yamanishi, K. Hyperbolic ordinal embedding. In *Proceedings of the 11th Asian Conference on Machine Learning*, pp. 1065–1080, 2019.

Tabaghi, P. and Dokmanic, I. Hyperbolic distance matrices. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1728–1738, 2020.

Tamuz, O., Liu, C., Belongie, S. J., Shamir, O., and Kalai, A. Adaptively learning the crowd kernel. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 673–680, 2011.

Terada, Y. and von Luxburg, U. Local ordinal embedding. In *Proceedings of the 31st International Conference on Machine Learning*, pp. 847–855, 2014.

Tifrea, A., Bécigneul, G., and Ganea, O. Poincare glove: Hyperbolic word embeddings. In *the 7th International Conference on Learning Representations*, 2019.

Tropp, J. A. An introduction to matrix concentration inequalities. *Foundations and Trends in Machine Learning*, 8(1-2):1–230, 2015.

van der Maaten, L. and Weinberger, K. Q. Stochastic triplet embedding. In *Proceedings of the 22nd IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, 2012.

Walter, J. A. and Ritter, H. J. On interactive visualization of high-dimensional data using the hyperbolic plane. In *Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 123–132, 2002.