Tight and fast generalization error bound of graph embedding in metric space

Atsushi Suzuki [1 2] **Atsushi Nitanda** [3 4] **Taiji Suzuki** [2 4] **Jing Wang** [5] **Feng Tian** [6] **Kenji Yamanishi** [2]

Abstract

Recent studies have experimentally shown that we can achieve in non-Euclidean metric space effective and efficient graph embedding, which aims to obtain the vertices' representations reflecting the graph's structure in the metric space. Specifically, graph embedding in hyperbolic space has experimentally succeeded in embedding graphs with hierarchical-tree structure, e.g., data in natural languages, social networks, and knowledge bases. However, recent theoretical analyses have shown a much higher upper bound on non-Euclidean graph embedding's generalization error than Euclidean one's, where a high generalization error indicates that the incompleteness and noise in the data can significantly damage learning performance. It implies that the existing bound cannot guarantee the success of graph embedding in non-Euclidean metric space in a practical training data size, which can prevent non-Euclidean graph embedding's application in real problems. This paper provides a novel upper bound of graph embedding's generalization error by evaluating the local Rademacher complexity of the model as a function set of the distances of representation couples. Our bound clarifies that the performance of graph embedding in non-Euclidean metric space, including hyperbolic space, is better than the existing upper bounds suggest. Specifically, our new upper bound is polynomial in the metric space's geometric radius R and can be $O(\frac{1}{S})$ at the fastest, where S is the training data size. Our bound is significantly tighter and faster than the existing one, which can be exponential in R and $O(\frac{1}{\sqrt{S}})$ at the fastest. Specific calculations on example cases show that graph embedding in non-Euclidean metric space can outperform that in Euclidean space with much smaller training data than the existing bound has suggested.

1. Introduction

Graphs are a fundamental form of real-world entities and their relations, such as words in natural languages, people in social networks, and objects in knowledge bases. Here, the vertices and edges of a graph correspond to the entities and the relations among them, respectively. Based on the formulation, ***graph embedding***, learning representations of the graph's vertices in a metric space has enabled numerous applications for those data, such as machine translation and sentiment analysis for natural language (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Tifrea et al., 2019), and community detection and link prediction for social network data (Hoff et al., 2002; Perozzi et al., 2014; Tang et al., 2015b;a; Grover & Leskovec, 2016), pathway prediction of biochemical network (Dale et al., 2010; MA Basher & Hallam, 2021), and link prediction and triplet classification for knowledge base (Nickel et al., 2011; Bordes et al., 2013; Riedel et al., 2013; Nickel et al., 2016; Trouillon et al., 2016; Ebisu & Ichise, 2018). The metric space where we get representations of the vertices is called the ***representation space*** in this paper. Graph embedding aims to obtain representations such that the metric reflects the relations defined by the edges. Specifically, we expect the representations of a couple of vertices to be close if they are connected and distant if not.

It is essential in the representation learning context to discuss a generic metric space, not only Euclidean space, as a representation space, although Euclidean space or the inner product space has been widely used (Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2017; Hoff et al., 2002; Perozzi et al., 2014; Tang et al., 2015b;a; Grover & Leskovec, 2016). It is because many studies have experimentally shown the effectivity or representation learning in non-Euclidean metric space, in particular, hyperbolic space (Nickel & Kiela, 2017; Ganea et al., 2018a; Sala et al., 2018; Ganea et al., 2018b; Chami et al., 2019; Gülçehre et al., 2019; Tifrea et al., 2019; Balazevic et al., 2019) since hyperbolic space can represent a graph with a hierarchical tree-like structure with arbitrarily small approximation error

[1] King's College London, UK [2] The Univerity of Tokyo, Japan [3] Kyushu Institute of Technology, Japan [4] RIKEN, Japan [5] University of Greenwich, UK [6] Duke Kunshan University, China. Correspondence to: Atsushi Suzuki <atsushi.suzuki.rd@gmail.com>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

(Gromov, 1987; Sarkar, 2011; Sala et al., 2018). This advantage comes from the property that the volume of hyperbolic space grows exponentially in its radius R. This is in contrast to Euclidean space, which has limitations in representing such a graph (Lamping & Rao, 1994; Ritter, 1999; Nickel & Kiela, 2017).

The above facts motivate us to use graph embedding in non-Euclidean space actively. However, the ability of some non-Euclidean space to represent a complex graph may lead to overfitting in graph embedding settings, where data are incomplete or noisy. It is because there is a trade-off between a model's representability and the potential of overfitting in general machine learning settings. Hence, in order for graph embedding users to select the best model, we need to evaluate each model's ***generalization error***, that is, how much the model's performance is badly influenced by incompleteness and noise of the data. Indeed, recent research (Suzuki et al., 2021a;b), for the first time, has provided upper bounds of representation learning in non-Euclidean space by converting the graph embedding problem to a linear discrimination analysis problem from Gramian matrices in the inner-product space or Minkowski space. Their results suggest that the generalization error of the representation learning's performance could be exponential in the radius R of the hyperbolic space that we use. This bound is in line with the volume of the space. Their evaluation implies that we might need an impractically large data size (e.g., $> 10^{72}$ as in Remark 11) to get a better performance graph embedding in hyperbolic space than in Euclidean space. Nevertheless, the following observations imply the existing bounds overestimate the generalization error.

- They do not consider the metric space's property. Even if the volume of a ball grows exponentially in its radius R as in hyperbolic space, the distance between two points in the ball is always smaller than $2R$. Hence, the generalization error might avoid an exponential dependency on the space's radius.

- They do not use the "local" model complexity around the optimal representations, resulting in a convergence rate $O(1/\sqrt{S})$ for data size S. According to past research (Bartlett et al., 2005; Koltchinskii, 2006) in the learning theory context, the generalization error can be $O(1/S)$ if the complexity of the "neighborhood" of the best hypothesis function is limited. In the graph embedding setting, the model is substantially finite-dimensional since there are finite representation couples only. Hence, it is highly possible that the "local" complexity is small enough.

Based on the above observation, we aim to derive a tighter and faster generalization error bound of graph embedding in metric space. The above observations imply that we have the potential to achieve a tighter and faster bound if we regard graph embedding's loss function as a function of the distance values of the finite representation couples. Indeed, we have achieved the aim by reformulating graph embedding's loss function as a restriction of the composition of a non-linear function and a linear function of the distances of pairs of representations. Specifically, our contributions are the following:

- We have derived a novel upper bound of the ***Rademacher complexity*** (Koltchinskii, 2001; Koltchinskii & Panchenko, 2000; Bartlett et al., 2002) of graph embedding's hypothesis function set and its local subset, called the ***local Rademacher complexity*** (Bartlett et al., 2005; Koltchinskii, 2006). The bound is tighter than existing ones for most cases since it is polynomial for the representation space's radius if the space is metric. The Rademacher complexity evaluation can apply to representation learning settings discussed in the past papers (Jain et al., 2016; Gao et al., 2018; Suzuki et al., 2021a;b) since their bounds were also derived from the Rademacher complexity evaluation.

- Based on the above global and local Rademacher complexity bound, we have derived a novel upper bound of graph embedding's generalization error. Our bound is tighter in that it is polynomial for the representation space's radius R if the space is metric and faster in that it is $O(1/S)$ at the fastest than the existing $O(1/\sqrt{S})$ bound.

- We have calculated specific bounds for graph embedding in Euclidean and hyperbolic spaces and derived a significantly improved upper bound of the data size that the graph embedding in hyperbolic space needs to outperform that in Euclidean space when the graph is a tree.

The remainder of the paper is organized as follows. Section 2 formulates the graph embedding in the learning theory style. Section 3 gives our assumptions and generalization error bounds, the main result of the paper. Section 4 provides examples of the application of our main result. Section 5 gives the core technical result to enable comparisons to previous work and discussions on potential future work. Section 6 compares our result with previous work based on Section 5. Section 7 discusses potential future work.

2. Preliminaries

Notation The symbol $:=$ indicates that its left side is defined by its right 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. For $D \in \mathbb{Z}_{>0}$, $\mathbb{R}^D$ denotes the set of D-dimensional real vectors. For $\boldsymbol{z} \in \mathbb{R}^D$, $\boldsymbol{z}^\top$ indicates its transpose. $\operatorname{sgn} : \mathbb{R} \to \{0, \pm 1\}$ is the sign function defined by $\operatorname{sgn}(r) = -1$ if $r < 0$, $\operatorname{sgn}(r) = +1$ if $r > 0$, and $\operatorname{sgn}(r) = 0$ if $r = 0$. For $r, r' \in \mathbb{R}$, we define $r \wedge r' := \min\{r, r'\}$ and $r \vee r' := \max\{r, r'\}$. For a finite set $\mathcal{V}$, we denote the number of elements in $\mathcal{V}$ by $|\mathcal{V}| \in \mathbb{Z}_{\geq 0}$, and we denote the set of two element subsets of $\mathcal{V}$ by $\mathrm{C}_{\mathcal{V}}$, i.e., $\mathrm{C}_{\mathcal{V}} \coloneqq \{\mathcal{A} \subset \mathcal{V} | |\mathcal{A}| = 2\}$. Note that $|\mathrm{C}_{\mathcal{V}}| = \frac{|\mathcal{V}|(|\mathcal{V}|-1)}{2}$ holds. For sets $\mathcal{A}$ and $\mathcal{B}$, we denote by $2^{\mathcal{A}}$ the power set on $\mathcal{A}$, and by $\mathcal{B}^{\mathcal{A}}$ the set of maps from $\mathcal{A}$ to $\mathcal{B}$. For example, $\mathbb{R}^{\mathcal{A}}$ denotes the set of real functions on $\mathcal{A}$. If $\mathcal{A}$ is a measurable space, we denote the set of measurable functions on $\mathcal{A}$ by $\mathcal{L}_0(\mathcal{A})$. We denote the expectation with respect to a random variable z that follows a distribution P by $\mathbb{E}_{z \sim \mathrm{P}}$.

2.1. True dissimilarity

First, we formulate the representation learning from pair-label couples, which is of interest in this paper. This includes graph embedding as a special case. Let $\mathcal{V}$ denote the entity set. We assume that there exists a ***true dissimilarity*** function $\Delta^* : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$, where $\Delta^*(u, v)$ indicates the true dissimilarity between entity u and entity v. The entities u and v are "similar" or strongly related if $\Delta^*(u, v)$ is small and "dissimilar" or weakly related if $\Delta^*(u, v)$ is large. Specifically, we fix a threshold $\theta \in \mathbb{R}$ and say u and v are similar if $\Delta^*(u, v) < \theta$ and dissimilar if $\Delta^*(u, v) > \theta$. Note that $\Delta^*(u, v) = \theta$ holds with probability at most zero in this paper, so we can ignore this corner case. Throughout this paper, we assume the symmetry of the dissimilarity function, i.e., $\Delta^*(u, v) = \Delta^*(v, u)$ for all $u, v \in \mathcal{V}$. We can regard the setting discussed in this section as graph embedding if there exists a true undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where $\mathcal{E} \subset \mathrm{C}_{\mathcal{V}}$, and the true dissimilarity is given by the distance function $\Delta_{\mathcal{G}}$ defined by the graph $\mathcal{G}$ and we set the threshold $\theta = 1.5$. Here, $\Delta_{\mathcal{G}}(u, v)$ is defined by the length of a shortest path in $\mathcal{G}$ between u and v. Note that $\Delta_{\mathcal{G}}(u, u) = 0$ for all $u \in \mathcal{V}$, and $\Delta_{\mathcal{G}}(u, u) = \infty$ if there exists no path between u and v. Here, $\Delta_{\mathcal{G}}(u, v) < \theta = 1.5$ if and only if $\{u, v\} \in \mathcal{E}$ or $u = v$. Thus, we can regard graph embedding as a special case of the discussion here.

Remark 1. We do not assume properties of the true dissimilarity function Δ^* other than the symmetry, though all dissimilarity function examples in this paper are metric.

Remark 2. Although we omit the case where $\Delta^*(u, v) = \theta$ for simplicity, considering this corner case is not difficult if we simply regard u and v are similar if $\Delta^*(u, v) = \theta$ and modifying following definitions slightly.

2.2. Representation space and the objective of representation learning

Fix some space $\mathcal{W}$ with a ***distance function*** $\Delta_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ that is symmetric, i.e., $\Delta_{\mathcal{W}}(w, w') = \Delta_{\mathcal{W}}(w', w)$ for all $w, w' \in \mathcal{W}$. Here, we consider two points $w, w' \in \mathcal{W}$ to be "distant" if $\Delta_{\mathcal{W}}(w, w')$ is large and "close" if $\Delta_{\mathcal{W}}(w, w')$ is small. We call $\mathcal{W}$ the ***representation space***. The most typical example is the D-dimensional Euclidean space $\left(\mathbb{R}^D, \Delta_{\mathbb{R}^D}\right)$, where $\Delta_{\mathbb{R}^D} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}_{\geq 0}$ defined by $\Delta_{\mathbb{R}^D}(\boldsymbol{z}, \boldsymbol{z}') = \sqrt{(\boldsymbol{z} - \boldsymbol{z}')^\top (\boldsymbol{z} - \boldsymbol{z}')}$. Note that our main theorem allows distance functions not satisfying non-negativity or triangle inequality. See Assumption 1 for the rigorous conditions.

The objective of representation learning is to get a map $\mathrm{w} : \mathcal{V} \to \mathcal{W}$ which maps an entity to a representation in $\mathcal{W}$, such that the representations are consistent to the true dissimilarity defined by Δ^*. Here, we call w the ***representation map***, and for $v \in \mathcal{V}$, we call $w_v \coloneqq \mathrm{w}(v) \in \mathcal{W}$ the ***representation*** of entity v. Specifically, the objective of representation learning is to find a good representation map w that satisfies

$$\Delta^*(u, v) \lesseqgtr \theta \Leftrightarrow \Delta_{\mathcal{W}}(w_u, w_v) \lesseqgtr \theta_{\mathcal{W}}, \tag{1}$$

for "most" $\{u, v\} \in \mathrm{C}_{\mathcal{V}}$. We quantify the meaning of "most" in Section 2.5. Here, $\theta_{\mathcal{W}} \in \mathbb{R}$ is a threshold value. To make the formulation compatible with learning theory's notation, we rewrite the above representation learning objective as follows. Define the true label function $y^* : \mathrm{C}_{\mathcal{V}} \to \{\pm 1\}$ by $y^*(\{u, v\}) \coloneqq \operatorname{sgn}(\Delta^*(u, v) - \theta)$. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a nondecreasing function and define the ***hypothesis function*** $f_{\mathrm{w},\psi} : \mathrm{C}_{\mathcal{V}} \to \mathbb{R}$ by

$$f_{\mathrm{w},\psi}(\{u, v\}) \coloneqq \psi(\Delta_{\mathcal{W}}(w_u, w_v)) - \psi(\theta_{\mathcal{W}}). \tag{2}$$

Then, we can see that (1) is equivalent to the following.

$$y^*(\{u, v\}) f_{\mathrm{w},\psi}(\{u, v\}) > 0. \tag{3}$$

Thus, our objective to find a representation map w that satisfies the above inequality for "most" $\{u, v\} \in \mathrm{C}_{\mathcal{V}}$.

Remark 3. We do not assume that the distance function $\Delta_{\mathcal{W}}$ is metric. Specifically, even if it violates the triangle inequality or has two points $w \neq w'$ such that $\Delta_{\mathcal{W}}(w, w') \leq 0$, our main theorem holds. Nevertheless, the triangle inequality is important to obtain a specific bound as discussed in Section 4.1.

Remark 4. Mathematically, we do not need the function ψ since we can achieve the same thing by simply composing the ψ to the original $\Delta_{\mathcal{W}}$ and create a new distance function $\Delta'_{\mathcal{W}}$. However, it often makes interpretations easier if we fix $\Delta_{\mathcal{W}}$ to the distance function of a metric space and vary ψ instead of varying $\Delta_{\mathcal{W}}$. One good example is our discussion in Section 6, where $\Delta_{\mathcal{W}}$ is fixed to the distance function of Euclidean or hyperbolic space, while ψ can vary.

2.3. Couple-label pair data and graph embedding

We have discussed the objective of representation learning in Section 2.2. To achieve the objective, we need to use some data that contain partial information about the true dissimilarity Δ^*. For simple discussion, this paper focus on representation learning using couple-label pair data, which includes graph embedding as an important special case. Still, our theory straightforwardly applies to existing settings, i.e., that in (Suzuki et al., 2021b) as we discuss in Section 5.

A couple-label pair data sequence is a sequence $(z_s)_{s=1}^{S}$ of pairs of an unordered entity couple and a label. Specifically, the s-th data point $z_s = (x_s, y_s)$ consists of a pair of an unordered entity couple $x_s = \{u_s, v_s\} \in \mathrm{C}_{\mathcal{V}}$ and a label $y_s \in \{\pm 1\}$. Here, $y_s = +1$ indicates that the s-th data point claims u_s and v_s being similar, i.e., $\Delta^*(u_s, v_s) < \theta$, and $y_s = -1$ indicates its converse, i.e., $\Delta^*(u_s, v_s) > \theta$. Nevertheless, this correspondence between the label y_s and dissimilarity $\Delta^*(u_s, v_s)$ does not always hold because the data point may be wrong owing to data noise. That is, the learning task is **agnostic** with the label noise. As discussed in Section 2.1, if the true dissimilarity is given by the distance function $\Delta_{\mathcal{G}}$ defined by the graph $\mathcal{G}$ and we set the threshold $\theta = 1.5$, then $y_s = +1$ claims that there exists an edge between u_s and v_s.

Remark 5. Strictly speaking, if the graph $\mathcal{G}$ is not connected $\Delta_{\mathcal{G}}$ can take $+\infty$, which our setting does not formally cover since the range of the true dissimilarity Δ^* is $\mathbb{R}$, not $\mathbb{R} \cup \{\pm\infty\}$. However, the extension of the range to $\mathbb{R} \cup \{\pm\infty\}$ is easy. In this paper, for notation simplicity, we limit the range to $\mathbb{R}$ and we only discuss connected graphs as examples.

2.4. Loss function

To obtain representations using data as we discussed in Section 2.3, we need to quantify how compatible representations are with the data. This is what a ***loss function*** does. This subsection defines the loss function for generic cases. The definitions in the remainder of this subsection consider a generic prediction setting from a feature space $\mathcal{X}$ to label space $\mathcal{Y}$ to compare the couple-label pair case to the general discussion later. Still, we can always specialize the discussion by substituting $\mathcal{X} = \mathrm{C}_{\mathcal{V}}$ and $\mathcal{Y} = \{\pm 1\}$. Given a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$, the loss of a hypothesis function $f : \mathcal{X} \to \mathbb{R}$ on a data point $(x_s, y_s) \in \mathcal{X} \times \mathcal{Y}$ is given by $\ell(x_s, y_s, f(x_s))$. In the couple-label pair case, our main interest on a data point (x_s, y_s) whether the hypothesis function's output $f(x_s)$ has the same sign as the label y_s has, as discussed in Section 2.2. That is, our interest is the sign of $y_s f(x_s)$. Hence, we mainly consider a ***margin-based loss***, that is, a loss function that can be defined by $\ell(x, y, t) \coloneqq \phi(yt)$, where $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is a non-increasing function. Here, we assume ϕ is non-increasing because it is desirable and deserve a low loss if $y_s f(x_s)$ is positive and vice versa. The function ϕ is called a ***representing function***. A typical example is the ***hinge loss function*** defined by $\phi_{\mathrm{hinge}}(t) \coloneqq \max\{-t + 1, 0\}$, which is non-increasing.

If the input of the loss function is unrestricted, the loss can be unbounded, which can lead to infinite risk. Hence, we introduce ***clipping*** following (Chapter 2, Steinwart & Christmann, 2008). For $M \in \mathbb{R}_{\geq 0}$, we define the ***clipped value*** $[t]_{-M}^{+M} \in [-M, +M]$ by

$$[t]_{-M}^{+M} \coloneqq \begin{cases} -M & \text{if } t \leq -M, \\ t & \text{if } t \in [-M, +M], \\ +M & \text{if } t \geq +M. \end{cases} \tag{4}$$

Fix $M \in \mathbb{R}_{\geq 0}$, and we say that a loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ is ***clippable*** at M if $\ell\left(x, y, [t]_{-M}^{+M}\right) \leq \ell(x, y, t)$ for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For example, the hinge loss function is a typical clippable loss.

2.5. Data distribution and risks

We assume that a data point is generated by a distribution P on $\mathcal{X} \times \mathcal{Y}$. Once a distribution P is given, our interest is the expectation of the loss of a hypothesis function f with respect to P. This expectation is called the ***expected risk*** of f with respect to the loss function ℓ and distribution P, denoted by $\mathscr{R}_{\ell,\mathrm{P}}(f)$. Here, the ***risk function*** $\mathscr{R}_{\ell,\mathrm{P}} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}_{\geq 0}$ is defined by

$$\mathscr{R}_{\ell,\mathrm{P}}(f) \coloneqq \mathbb{E}_{(x,y)\sim\mathrm{P}} h_{\ell,f}(x, y), \tag{5}$$

where $h_{\ell,f} : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}_{\geq 0}$ is defined by $h_{\ell,f}(x, y) \coloneqq \ell(x, y, f(x))$. We also define the ***clipped expected risk*** $[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) \coloneqq \mathbb{E}_{(x,y)\sim\mathrm{P}}[h_{\ell,f}]_{-M}^{+M}(x, y)$, where $[h_{\ell,f}]_{-M}^{+M}(x, y) \coloneqq \ell\left(x, y, [f(x)]_{-M}^{+M}\right)$. Now, we can formally state that the objective of representation learning is to minimize the expected risk. Since the definition of the risk involves expectation, we only consider a measurable function as a hypothesis function, i.e., $f \in \mathcal{L}_0(\mathcal{X})$. However, in the couple-pair label case, since $\mathcal{X}$ is a finite set and we consider discrete topology, we have that $\mathcal{L}_0(\mathcal{X}) = 2^{\mathcal{X}}$. Thus, every function on $\mathcal{X}$ is measurable and we can ignore the discussion on measurability. Although the objective of representation learning is to minimize the expected risk, we cannot directly do that since we cannot directly observe the data distribution. Instead, since we have a data sequence $(x_s, y_s)_{s=1}^{S}$, we minimize the ***empirical risk***

$$\mathscr{R}_{\ell,\mathrm{S}}(f) = \mathbb{E}_{(x,y)\sim\mathrm{S}} h_{\ell,f}(x, y) = \frac{1}{S} \sum_{s=1}^{S} h_{\ell,f}(x_s, y_s). \tag{6}$$

which is the risk calculated on the empirical measure $\mathrm{S} : 2^{\mathcal{X}\times\mathcal{Y}} \to \mathbb{R}$ defined by $\mathrm{S} \coloneqq \frac{1}{S} \sum_{s=1}^{S} \delta_{(x_s, y_s)}$. Here, $\delta_{(x,y)}(\mathcal{A}) = 1$ if $(x, y) \in \mathcal{A}$ and $\delta_{(x,y)}(\mathcal{A}) = 0$ otherwise.

Or, we might minimize the clipped version $[\mathscr{R}_{\ell,\mathrm{S}}]_{-M}^{+M}$. We remark that if the loss function ℓ is clippable, then $[\mathscr{R}_{\ell,\mathrm{S}}]_{-M}^{+M} \leq \mathscr{R}_{\ell,\mathrm{S}}(f)$ for all $f \in \mathcal{L}_0(\mathcal{X})$. Following (Steinwart & Christmann, 2008), we define empirical risk minimization below so that the definition includes minimization of both versions.

Definition 1. Let $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$, and fix $\epsilon \in \mathbb{R}_{\geq 0}$. Then a map $\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^S \to \mathcal{F}$ is called an ***ϵ-approximation clipped empirical risk minimization*** (***ϵ-CERM***) if it satisfies

$$[\mathscr{R}]^{+M}_{-M\,\ell,\mathrm{S}}\left(\mathfrak{A}\left((z_s)_{s=1}^{S}\right)\right) \leq \inf\left\{\mathscr{R}_{\ell,\mathrm{S}}(f) | f \in \mathcal{F}\right\} + \epsilon, \tag{7}$$

for all $(z_s)_{s=1}^{S} \in (\mathcal{X} \times \mathcal{Y})^S$ and empirical measure S determined by $(z_s)_{s=1}^{S}$. A 0-CERM is called a ***clipped empirical risk minimization*** (***CERM***).

Remark 6. The left hand side of the inequality in (7) is not clipped. Hence, if the loss function ℓ is clippable and a map $\mathfrak{A}$ minimizes either the non-clipped empirical risk or the clipped one, then it is a CERM.

Since we want to have as low a risk as possible, we are interested in the infimum of the risk. We denote the infimum of the risk in a given hypothesis function set $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$ by $[\mathscr{R}^{*,\mathcal{F}}_{\ell,\mathrm{P}}]^{+M}_{-M}$, defined by $[\mathscr{R}^{*,\mathcal{F}}_{\ell,\mathrm{P}}]^{+M}_{-M} := \inf\left\{[\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f) \middle| f \in \mathcal{F}\right\}$.

The infimum $[\mathscr{R}^{*}_{\ell,\mathrm{P}}]^{+M}_{-M} := [\mathscr{R}^{*,\mathcal{L}_0(\mathcal{X})}_{\ell,\mathrm{P}}]^{+M}_{-M}$ of the expected risk over all hypothesis functions is called the ***Bayes risk***.

Since we try to achieve the Bayes risk using a CERM, we are interested in how well it goes. Hence, we will evaluate the ***excess risk*** defined by $[\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f) - [\mathscr{R}^{*}_{\ell,\mathrm{P}}]^{+M}_{-M}$, where f is a CERM. We can regard the excess risk of a CERM as a quantification of the generalization error.

3. Fast rate of generalization error bound in representation learning

This section states our upper bounds of the excess risk in representation learning on couple-label pair data.

Assumption 1. Fix $M \in \mathbb{R}_{>0}$. Consider the following conditions regarding the representation space $\mathcal{W}$, the dissimilarity function $\Delta_{\mathcal{W}}$, the function ψ, the loss function ℓ, and the data distribution P.

(C1) The random variables $z_1, z_2, \ldots, z_S$ follow the distribution P on $\mathcal{X} \times \mathcal{Y}$ mutually independently.

(C2) The representation space $\mathcal{W}$ is a topological space and **compact**.

(C3) The dissimilarity function $\Delta_{\mathcal{W}} : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ is a **continuous** symmetric function on $\mathcal{W} \times \mathcal{W}$.

(C4) The feature and label spaces are $\mathcal{X} = \mathrm{C}_{\mathcal{V}}$ and $\mathcal{Y} = \{\pm 1\}$, and the hypothesis function set $\mathcal{F}$ is given by $\mathcal{F} = \mathcal{F}_{\mathrm{w},\psi} := \{f_{\mathrm{w},\psi} | \mathrm{w} : \mathcal{V} \to \mathcal{W}\}$, where $f_{\mathrm{w},\psi}$ is defined by (2) and $\psi : \mathbb{R} \to \mathbb{R}$ is a **continuous** non-decreasing function.

(C5) The loss function ℓ is clippable at M.

(C6) The loss function ℓ is margin-based with a representing function $\phi : \mathbb{R} \to \mathbb{R}_{\geq 0}$.

(C7) The loss function ℓ satisfies the ***supremum bound condition***, i.e., $\exists B \in \mathbb{R}_{>0}, \forall (x, y) \in \mathcal{X} \times \mathcal{Y}, \forall t \in [-M, M] : \ell(x, y, t) \leq B$.

(C8) The representing function ϕ is Lipschitz continuous i.e., there exists a constant $L \in \mathbb{R}_{\geq 0}$ such that $\phi(t - t') \leq L|t - t'|$ for any $t, t' \in [-M, M]$.

(C9) There exists a ***Bayes decision function*** in $\mathcal{F}$, i.e., there exists a hypothesis function $f^* \in \mathcal{F}$ that satisfies $[\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f^*) = \inf\left\{[\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f) \middle| f \in \mathcal{L}_0(\mathcal{X})\right\}$.

(C10) There exists $\vartheta \in [0, 1]$ such that the ***variance bound condition*** holds, i.e., there exists $U \in \mathbb{R}_{\geq 0}$ such that for all $f \in \mathcal{F}$,

$$\begin{aligned} &\mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}}\left[[f(x)]^{+M}_{-M} - [f^*(x)]^{+M}_{-M}\right]^2 \\ &\leq U\left[[\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}]^{+M}_{-M}(f^*)\right]^{\vartheta}. \end{aligned} \tag{8}$$

Remark 7. In Assumption 1,

(a) If $\mathcal{W}$ is a metric space, then the condition (C2) holds if and only if $\mathcal{W}$ is totally bounded and complete. Also, its distance function always satisfy (C3). For example, a closed ball in finite-dimensional Euclidean or hyperbolic space can satisfy (C2) and (C3). Furthermore, if $\mathcal{W}$ is a subset of finite-dimensional Euclidean space, then (C2) holds if and only if $\mathcal{W}$ is bounded and closed. If $\mathcal{W}$ is a subset of finite-dimensional inner-product space, then the (negative) inner-product function is a continuous symmetric function, which satisfies (C3) as a distance function.

(b) In (C10), $\vartheta = 1$ requires the "strong-convexity" of the loss function ℓ with respect to the hypothesis function f, which is assumed in, e.g., (Bartlett et al., 2005; Koltchinskii, 2006).

(c) The conditions (C8) and (C10) imply the following

$$\begin{aligned} &\mathbb{E}_{(x,y) \sim \mathrm{P}}\left[[h_{\ell,f}]^{+M}_{-M}(x, y) - [h_{\ell,f^*}]^{+M}_{-M}(x, y)\right]^2 \\ &\leq V\left[\mathscr{R}_{\ell,\mathrm{P}}{}^{M}(f) - \mathscr{R}_{\ell,\mathrm{P}}{}^{M}(f^*)\right]^{\vartheta}, \end{aligned} \tag{9}$$

for $f \in \mathcal{F}$ with $V = L^2 U$, which corresponds to the condition assumed in (Section 7, Steinwart & Christmann, 2008).

We will discuss specific examples satisfying Assumption 1 in Section 4. The following is our main result.

Theorem 1. *Suppose that (C1) to (C8) in Assumption 1 holds. Let $L, B \in \mathbb{R}_{\geq 0}$ be constants that satisfy the inequalities in items (C7) and (C8) in Assumption 1, respectively, and define $F \in \mathbb{R}_{\geq 0}$ by*

$$F^2 = \max_{f \in \mathcal{F}} \sum_{\{u,v\} \in \mathrm{C}_{\mathcal{V}}} [f(\{u,v\})]^2. \tag{10}$$

Fix $\delta \in \mathbb{R}_{>0}$ and $\epsilon \in \mathbb{R}_{\geq 0}$. Then

(i) There exists a measurable (0-)CERM.

(ii) Any ϵ-CERM $\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^S \to \mathbb{R}$ satisfies

$$\begin{aligned}&[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}\Big(\mathfrak{A}\Big((z_s)_{s=1}^S\Big)\Big) - \inf\{\mathscr{R}_{\ell,\mathrm{P}}(f) | f \in \mathcal{F}\} \\ &\leq r_0(S) + \beta'(S) + \epsilon,\end{aligned} \tag{11}$$

in probability at least $1 - \delta$, where

$$r_0(S) \coloneqq 4LF \cdot \left(\frac{2}{S}\right)^{\frac{1}{2}}, \beta'(S) \coloneqq B_0 \cdot \left(\frac{\ln \frac{1}{\delta}}{S}\right)^{\frac{1}{2}}. \tag{12}$$

(iii) In addition, suppose items (C9) and (C10) hold, and let $U \in \mathbb{R}_{\geq 0}$ and $\vartheta \in [0, 1]$ be constants that satisfy the inequalities in item (C10) of Assumption 1 and fix $B_0 > B$. Then every ϵ-CERM $\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^S \to \mathbb{R}$ satisfies

$$\begin{aligned}&[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}\Big(\mathfrak{A}\Big((z_s)_{s=1}^S\Big)\Big) - [\mathscr{R}_{\ell,\mathrm{P}}^{*}]_{-M}^{+M} \\ &\leq \min\{r_i(S) | 0 \leq i \leq |\mathrm{C}_{\mathcal{V}}|\} \vee \alpha(S) \vee \beta(S) + 3\epsilon,\end{aligned} \tag{13}$$

in probability at least $1 - \delta$. Here, $\alpha, \beta : \mathbb{Z}_{>0} \to \mathbb{R}_{\geq 0}$ are defined by

$$\alpha(S) \coloneqq 3 \cdot \left(\frac{72\big(B^{2-\vartheta} \vee L^2 U\big) \ln \frac{3}{\delta}}{S}\right)^{\frac{1}{2-\vartheta}}, \beta(S) \coloneqq \frac{15 B_0}{S} \ln \frac{3}{\delta}, \tag{14}$$

and, $r_i(S)$ for $1, \ldots |\mathrm{C}_{\mathcal{V}}|$ is defined as the only positive solution of the equation $r = {}^{30\chi_i(r)}/\sqrt{S}$ for r, where

$$\chi_i(r) \coloneqq 2L\sqrt{2F^2\left(\frac{Ur^{\vartheta}}{4F^2} i + \mu_{\mathrm{P}_{\mathcal{X}}}(|\mathrm{C}_{\mathcal{V}}| - i)\right)}, \tag{15}$$

with $\mu_{\mathrm{P}_{\mathcal{X}}}(j) \coloneqq \min_{\mathcal{D} \subset \mathrm{C}_{\mathcal{V}}, |\mathcal{D}| = j} \mathrm{P}_{\mathcal{X}}(\mathcal{D})$. In particular,

$$r_{|\mathrm{C}_{\mathcal{V}}|}(S) \coloneqq 3 \cdot \left(1800 \cdot |\mathrm{C}_{\mathcal{V}}| \cdot \frac{L^2 U}{S}\right)^{\frac{1}{2-\vartheta}}. \tag{16}$$

We define $r_i(S) = 0$ if $LFU = 0$.

Remark 8. In Theorem 1,

(a) Although Assumption 1 does not explicitly assume the finiteness of F, it follows from the conditions (C2), (C3), and (C4) since F is defined as the maximum of a continuous function from a compact space $\mathcal{W}^{|\mathcal{V}|}$.

(b) The bound (13) is $O\Big(\big(\frac{1}{S}\big)^{\frac{1}{(2-\vartheta)}}\Big)$, which is faster than $O\Big(\big(\frac{1}{S}\big)^{\frac{1}{2}}\Big)$ if $\vartheta > 0$. In particular, it is $O\big(\frac{1}{S}\big)$ if $\vartheta = 1$.

(c) We have that $\min\{r_i(S) | 0 \leq i \leq |\mathrm{C}_{\mathcal{V}}|\} \leq r_0(S) \wedge r_{|\mathrm{C}_{\mathcal{V}}|}(S)$, whose right hand side is always analytically obtained. Here, $r_0(S) \gtreqless r_{|\mathrm{C}_{\mathcal{V}}|}(S) \Leftrightarrow S \gtreqless 7200 L^2 F^2 \left(\frac{U|\mathrm{C}_{\mathcal{V}}|}{4F^2}\right)^{\frac{2}{\vartheta}}$ if $\vartheta > 0$. This implies that the additional conditions in (iii) provides a faster rate for large S if $\vartheta > 0$. Note that we can ignore α and β unless we consider exponentially small δ. It is because α and β are in no slower order with respect to S than $r_{|\mathrm{C}_{\mathcal{V}}|}$ and r_0, respectively, and $r_{|\mathrm{C}_{\mathcal{V}}|}$ and r_0 depend on $|\mathcal{V}|$ and F while α and β are independent of them.

(d) As we can see from the definition of χ_i, the behavior of $r_i(S)$ for $i = 1, 2, \ldots, |\mathrm{C}_{\mathcal{V}}| - 1$ depend on $\mathrm{P}_{\mathcal{X}}$. If $\mathrm{P}_{\mathcal{X}}$ is the uniform distribution on $\mathcal{X}$, then $r_i(S) \geq r_0 \wedge r_{|\mathrm{C}_{\mathcal{V}}|}$ for $i = 1, 2, \ldots, |\mathrm{C}_{\mathcal{V}}| - 1$. Hence, we cannot improve the bound from $r_0(S) \wedge r_{|\mathrm{C}_{\mathcal{V}}|}(S)$. As an extreme example of the other direction, consider the case where there exists some $\mathcal{D} \subset \mathrm{C}_{\mathcal{V}}$ satisfies $\mathrm{P}_{\mathcal{X}}(\mathcal{D}) = 1$. Then we have that $r_{|\mathcal{D}|} \coloneqq 3 \cdot \left(1800 \cdot |\mathcal{D}| \cdot \frac{L^2 U}{S}\right)^{\frac{1}{2-\vartheta}}$. This is given by replacing $|\mathrm{C}_{\mathcal{V}}|$ in $r_{|\mathrm{C}_{\mathcal{V}}|}$ with $|\mathcal{D}|$. In particular, $r_{|\mathcal{D}|} \leq r_{|\mathrm{C}_{\mathcal{V}}|}$ and the equality holds if and only if $\mathcal{D} = \mathrm{C}_{\mathcal{V}}$. This result is natural since $\mathrm{P}_{\mathcal{X}}(\mathcal{D}) = 1$ means that we can ignore $\mathrm{C}_{\mathcal{V}} \setminus \mathcal{D}$.

Specific advantages of Theorem 1 over existing results will be discussed in Section 6.

4. Examples

Assumption 1 and Theorem 1 are given in a general form, including many parameters such as U, q, and F, which depend on the situation. This section gives specific examples of calculating these values in some application cases, and a comparison between Euclidean and hyperbolic spaces using the calculations.

4.1. Representation space and F

We assume that $\psi(t) = t^{\tau}$ for $\tau \geq 1$, as a simplest case. If $(\mathcal{W}, \Delta_{\mathcal{W}})$ is a metric space, whose radius is R, then from the triangle inequality, we have that $F^2 \leq |\mathrm{C}_{\mathcal{V}}|\big((2R)^{\tau} - (\theta_{\mathcal{W}})^{\tau}\big)^2 \leq |\mathrm{C}_{\mathcal{V}}|(2R)^{2\tau}$ if $\theta_{\mathcal{W}} \in [0, 2R]$. This is the worst case, and we have the following better bound for Euclidean space.

Lemma 2. *If $\mathcal{W}$ is a subset of a closed ball with radius $2R$ in Euclidean space, then $F^2 \leq \frac{|\mathcal{V}|}{8}(2R)^{2\tau}$.*

Here, the right side is linear for $|\mathcal{V}|$. On the other hand, the following lemma states that hyperbolic space almost achieves the worst case if the diameter is sufficiently large.

Lemma 3. *If $\mathcal{W}$ is a closed ball of radius R in hyperbolic space (dimension $D \geq 2$), then $\frac{F^2}{(2R)^{2\tau}} \to |\mathrm{C}_{\mathcal{V}}|$.*

The above result, at one glance, suggests that a Euclidean ball is better than a hyperbolic ball. However, the discussion is not trivial since hyperbolic space usually has a better approximation error. We will compare in Section 6 a Euclidean ball and a hyperbolic ball, considering both the approximation and generalization errors.

4.2. Hinge loss and ϑ

The upper bound in Theorem 1 heavily depends on ϑ. The value ϑ is determined by P and ℓ, but its calculation is not trivial. As an example case, we introduce the hinge loss case since it has been widely used as the loss function of the support vector machine (Cortes & Vapnik, 1995) and mainly discussed in the context of generalization error analysis in the classification problem (Steinwart & Christmann, 2008; Jain et al., 2016; Gao et al., 2018; Suzuki et al., 2021a;b). Suppose the loss function is the hinge loss, i.e., $\phi(t) = \max\{-t+1, 0\}$. Then, it is known that the parameter ϑ of the variance bound condition (C10) in Assumption 1 depends on the data distribution. Define $\eta : \mathcal{X} \to [0,1]$ by $\eta(x) := \frac{\mathrm{P}(\{(x,+1)\})}{\mathrm{P}_{\mathcal{X}}(\{x\})}$. Note that we can ignore the definition for x such that $\mathrm{P}_{\mathcal{X}}(\{x\}) = 0$ since it is about a measure-zero space. We say that a distribution P on $x \times \{\pm 1\}$ has ***noise exponent*** $q \in \mathbb{R}_{\geq 0}$ with constant $c \in \mathbb{R}_{>0}$ if $\mathrm{P}_{\mathcal{X}}(\{x \in \mathcal{X} \,|\, |2\eta(x) - 1| < t\}) \leq (ct)^q$, and noise exponent $+\infty$ with constant $c \in \mathbb{R}_{>0}$ if $\mathrm{P}_{\mathcal{X}}(\{x \in \mathcal{X} \,|\, |2\eta(x) - 1| < {}^{3}\!/\!{}_{c}\}) = 0$, where $\mathrm{P}_{\mathcal{X}}$ is the marginal distribution of P on $\mathcal{X}$ defined by $\mathrm{P}_{\mathcal{X}}(\mathcal{A}) := \mathrm{P}(\mathcal{A} \times \{\pm 1\})$ for a measurable set $\mathcal{A} \subset \mathcal{X}$. Here, a large q indicates a small noise. The condition $q = \infty$ corresponds to the strong low-noise condition, which has been assumed in, e.g., (Koltchinskii & Beznosova, 2005). We have the following, using existing results about ϑ for the hinge loss (e.g., Chapter 8, Steinwart & Christmann, 2008).

Corollary 4. *Suppose that conditions (C1) to (C4) in Assumption 1 are satisfied and the loss function is the hinge loss given by $\phi(t') := \max\{1 - t', 0\}$. Define F by (10) and let $M = 1$. Fix $\delta \in \mathbb{R}_{>0}$ and $\epsilon \in \mathbb{R}_{\geq 0}$. Then (i) there exists a measurable (0-)CERM. (ii) With $L = 1$, (ii) of Theorem 1 holds. (iii) In addition, if the condition (C9) holds and the distribution* P *has noise exponent $q \in \mathbb{R}_{\geq 0}$ with constant $c \in \mathbb{R}_{>0}$, then, with $B_0 > 2$, $\vartheta = \frac{q}{q+1}$, and $U = 6c^{\frac{q}{q+1}}$, (iii) of Theorem 1 holds.*

4.3. Improved comparison: Euclidean vs hyperbolic.

Suzuki et al. (2021b) showed a sufficient condition with respect to S for graph embedding in hyperbolic space to be better than that in Euclidean space. Following their paper's setting, we give a sufficient condition based on Theorem 1.

Assume that the posterior distribution is given by $\eta(\{u,v\}) = \frac{1}{2}(1 + \xi y^*(\{u,v\}))$, where $\xi \in [0,1]$. Fix ψ. The hypothesis function $f_{\mathrm{w},\psi}$ given by a representation map w gives a Bayes decision function if and only if

$$\begin{aligned} &y^*(\{u,v\}) = +1 \Rightarrow \psi(\Delta_{\mathcal{W}}(w_u, w_v)) \leq \psi(\theta_{\mathcal{W}}) - 1, \\ &y^*(\{u,v\}) = -1 \Rightarrow \psi(\Delta_{\mathcal{W}}(w_u, w_v)) \geq \psi(\theta_{\mathcal{W}}) + 1. \end{aligned} \tag{17}$$

Note that the above condition is stronger than (1).

In the following, $\mathcal{R}^D$ and $\mathcal{H}^D$ denote the D-dimensional Euclidean space and hyperbolic space, respectively. For a representation space $\mathcal{W}$ and a representation map w, we define $v_{\mathcal{W}} : \mathcal{W}^{\mathcal{V}} \to \mathbb{Z}_{\geq 0}$ by $v_{\mathcal{W}}(\mathrm{w}) := |\{\mathcal{D} | \forall\{u,v\} \in \mathcal{D} : \{u,v\} \text{ violates } (17)\}|$ and $v_{\min}(\mathcal{W}) := \min\{v_{\mathcal{W}}(\mathrm{w}) | \mathrm{w} : \mathcal{W} \to \mathcal{V}\}$. Then, we can see that $\mathscr{R}^{*,\mathcal{F}_{\mathcal{W},\psi}}_{\ell,\mathrm{P}} - \mathscr{R}^{*}_{\ell,\mathrm{P}} = \frac{v_{\min}(\mathcal{W})}{|\mathrm{C}_{\mathcal{V}}|}\xi$. If $\mathcal{W}$ is a metric space, let a closed ball with radius R in $\mathcal{W}$ denoted by $\mathcal{B}[R; \mathcal{W}]$.

If the true dissimilarity Δ^* is the graph distance of a tree, the following lemmata regarding $v_{\min}\big(\mathcal{B}[R; \mathcal{R}^2]\big)$ and $v_{\min}\big(\mathcal{B}[R; \mathcal{H}^2]\big)$ hold as straightforward modifications of results in (Sarkar, 2011; Suzuki et al., 2019; 2021a). (See the supplementary materials for the proofs).

Lemma 5. *Suppose that $(\mathcal{V}, \mathcal{E})$ is a tree and $\Delta^* : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ is given by its graph distance. Then, there exist $R \in \mathbb{R}_{\geq 0}$ such that $v_{\mathit{min}}\big(\mathcal{B}[R; \mathcal{H}^D]\big) = 0$ for any D.*

Lemma 6. *Let $p(D)$ be the packing number of the D-dimensional unit sphere with the unit distance. In particular, $p(2) = 5$. Suppose that the true dissimilarity Δ^* is given by the graph distance of a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. Then, $v_{\mathit{min}}\big(\mathcal{B}\big[R; \mathcal{R}^2\big]\big)$ is larger than or equal to the number of disjoint $(p(D)+1)$-star subgraphs in the graph.*

Lemmata 5 and 6 help the comparison between embedding in Euclidean space and hyperbolic space. The following is an example of a specific comparison in the setting discussed in (Suzuki et al., 2021b). For a more general discussion, see Appendix F.

Example 1. We consider the complete balanced λ-ary tree with height h, and the noise margin $\xi = \frac{1}{2}$. Suppose $\lambda = 5$ and $h = 4$. Here, we have that $|\mathcal{V}| = 156$ and Lemma 5 gives $R = 39.51$. If $S \geq 1.19 \times 10^9$ for $\tau = 1$ or $S \geq 7.43 \times 10^{12}$ for $\tau = 2$, then in probability at least $1 - 2^{-10}$, the expected risk of a CERM using $\mathcal{B}[R; \mathcal{H}^2]$ is better than that using any ball in $\mathcal{R}^2$.

Remark 9. The above evaluation uses the approximation error of the embedding using $\mathcal{R}^2$ as the lower bound of the error by ERM. We may obtain a better threshold in the near future once we obtain a good lower bound of the generalization error of representation learning using $\mathcal{R}^2$.

4.4. ℓ^1 embedding

Theorem 1 can apply to embedding using a general metric space. Here, we discuss an important example, ℓ^1 embedding in the same setting as in Section 6. The ℓ^1 embedding is representation learning in (a subset of) the metric space $\left(\mathbb{R}^D, \Delta_{D,1}\right)$, where $\Delta_{D,1}$ is the Manhattan distance $\Delta_{D,1}(\boldsymbol{w}, \boldsymbol{w}') = \left\|\boldsymbol{w} - \boldsymbol{w}'\right\|_1$. Here, $\left\|\cdot\right\|_1$ is the 1-norm operator. The following lemma holds for ℓ^1 embedding.

Lemma 7. *If $\mathcal{W}$ is a subset of a closed ball with radius R in a D-dimensional Manhattan distance space $\left(\mathbb{R}^D, \Delta_{D,1}\right)$, then we have that*

$$F^2 \leq \frac{|\mathcal{V}| D^\tau}{8} (2R)^{2\tau}. \tag{18}$$

The proof of Lemma 7 is easy from Lemma 2, the inequality $\Delta_{D,1}(\boldsymbol{w}, \boldsymbol{w}') \leq \sqrt{D} \Delta_{\mathbb{R}^D}(\boldsymbol{w}, \boldsymbol{w}')$, and the fact that, for any Manhattan distance ball with radius R, there always exists a Euclidean ball with radius R that covers it (since $\Delta_{D,1}(\boldsymbol{w}, \boldsymbol{w}') \geq \Delta_{\mathbb{R}^D}(\boldsymbol{w}, \boldsymbol{w}')$).

One interesting suggestion here is that it depends on the dimension. This is different from our bound on Euclidean space or hyperbolic space. The dependency of the generalization error bound of ℓ^1-embedding on the space dimension has a significant meaning in discussing the approximation-generalization (or bias-variance) trade-off of ℓ^1-embedding since we know specific results on the dependency of ℓ^1-embedding's approximation error on the space dimension D, for example:

Proposition 8 (A corollary from Proposition 11.1.4 in (Deza et al., 1997))**.** *The following statements are equivalent to each other:*

- *A tree T has $2D$ leaves.*
- *A tree T can be embedded with zero distortion in the D-dimensional Manhattan distance space.*

Hence, Proposition 8 and Lemma 7 quantitatively state a trade-off: if we increase D then the approximation error decreases but the generalization error increases. Interestingly, if we try to achieve the zero approximation error by setting $D = \frac{|\mathcal{V}|}{2}$, then the generalization error is the same as that of hyperbolic space given in Lemma 3, up to a constant factor. The above result indicates that we suffer from almost the same generalization error if we try to achieve an arbitrarily small approximation error for a tree, whether we use a hyperbolic space or a Manhattan space.

5. Core evaluation: Rademacher complexity

In this section, we provide the core technical result used to prove Theorem 1, to make an essential comparison in Section 6 between our results and existing results, without being influenced by the loss function's non-essential difference.

Our proof depends on the standard schemes in the statistical learning theory using the ***Rademacher complexity (RC)***.

Definition 2. Let $\sigma_1, \sigma_2, \dots, \sigma_S, z_1, z_2, \dots, z_S$ be mutually independent random variables, where each of $\sigma_1, \sigma_2, \dots, \sigma_S$ takes values $\{-1, +1\}$ with equal probability and each of $z_1, z_2, \dots, z_S$ follows some distribution P on a set $\mathcal{Z}$. The Rademacher complexity (RC) $\mathrm{Rad}_{\mathrm{P},S}(\mathcal{F})$ of a function set $\mathcal{F} \subset \mathcal{L}_0(\mathcal{Z})$ on P is defined by

$$\mathrm{Rad}_{\mathrm{P},S}(\mathcal{F}) \coloneqq \mathbb{E}_{(z_s)_{s=1}^S} \mathbb{E}_{(\sigma_s)_{s=1}^S} \left[\frac{1}{S} \sup_{f \in \mathcal{F}} \sum_{s=1}^{S} \sigma_s f(z_s) \right]. \tag{19}$$

In the following, we fix a measurable loss function $\ell : \mathcal{X} \times \mathcal{Y} \times \mathbb{R} \to \mathbb{R}_{\geq 0}$ and hypothesis function set $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$, and we define $h_{\ell,\mathcal{F}} \subset \mathcal{L}_0(\mathcal{X} \times \mathcal{Y})$ by $h_{\ell,\mathcal{F}} \coloneqq \{h_{\ell,f} | f \in \mathcal{F}\}$ and local hypothesis function set $\mathcal{F}_r \coloneqq \left\{ f \in \mathcal{F} \middle| [\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \leq r \right\}$ for $r \in \mathbb{R}_{\geq 0}$. Existing research, (e.g., Bartlett & Mendelson, 2002), has shown that we can obtain an upper bound of generalization error proportional to $\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}})$. It is also shown (e.g., Bartlett et al., 2005; Koltchinskii, 2006) that we can obtain a faster upper bound by evaluating $\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}_r})$, which we call the ***local Rademacher complexity (LRC)***. For the above reason, we are interested in the RC and LRC. Our evaluation of the RC and LRC in the couple-label data learning setting is the following.

Theorem 9. *Assume that the conditions (C2) to (C4), (C6) and (C8) hold. Then, we have that*

$$\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}_r}) \leq \min_{i=0,1,\dots,|\mathrm{C}_\mathcal{V}|} \frac{\chi_i(r)}{\sqrt{S}}, \tag{20}$$

where χ_i is defined by (15). In particular, by substituting $r = +\infty$, we have that $\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}}) = \mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}_\infty}) = 2LF \cdot \left(\frac{2}{S}\right)^{\frac{1}{2}}$.

Remark 10. Regarding Theorem 9,

(a) To the best of our knowledge, Theorem 9 is the first LRC evaluation in the context of representation learning, including couple-label pair data learning and graph embedding.

(b) Theorem 9 implies that we can have a meaningful LRC evaluation even without regularization, though we need it for e.g., the support vector machine analysis (Steinwart & Christmann, 2008). It is advantageous since it can exploit the non-Euclidean space's representability in the resulting upper bound.

(c) We can straightforwardly update existing generalization error bounds of graph embedding based on the RC, such as that in (Suzuki et al., 2021b), using the above bound, although this paper's discussion focuses on our simplest couple-label pair setting to save space.

As explained, the RC evaluation is substantial in deriving generalization error bounds, regardless of the specific form of the loss function. The discussion makes us ready for comparison in Section 6 between existing results and ours.

6. Related work and comparison

The generalization error of representation learning has been studied for the ordinal data case (Jain et al., 2016; Suzuki et al., 2021a) and where random variables associated with entities are observed (Wang et al., 2018). Still, the first paper that has derived a generalization error bound for a typical graph embedding setting is (Gao et al., 2018), although this paper only considers linear space and gives a result with some unevaluated term. To the best of our knowledge, only (Suzuki et al., 2021b) considers the generalization error for graph embedding in non-Euclidean space, including hyperbolic space. This section aims to compare our result with the result by (Suzuki et al., 2021b). In (Suzuki et al., 2021b), the positive-negative example data case is mainly discussed, which needs a large space to introduce and has a loss function different from ours. However, since the core technique of their result is also the RC evaluation, we can make an essential comparison between them throughout the evaluations. The following is the result by (Suzuki et al., 2021b).

Corollary 10 (Rademacher complexity evaluation by (Suzuki et al., 2021b))**.** *Let $\mathcal{W}$ be a closed ball with radius R in D-dimensional Euclidean space $\mathcal{R}^D$ or hyperbolic space $\mathcal{H}^D$. Let $\psi(t) = \gamma(t^2)$ for Euclidean case and $\psi(t) = \gamma(\cosh t)$ for hyperbolic space case, where $\gamma : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is a non-descreasing Lipschitz continuous function whose Lipschitz constant is L_γ. Also, let the loss function be the hinge loss given by $\phi(t') := \max\{1 - t', 0\}$. Then*

$$\mathrm{Rad}_{\mathrm{P},S}\big(h_{\ell,\mathcal{F}_{\mathcal{W},\psi}}\big) \leq \frac{\omega(R)}{S} L_\gamma |\mathcal{V}| \Big(\sqrt{2S\nu \ln |\mathcal{V}|} + \frac{\kappa}{3} \ln |\mathcal{V}|\Big), \tag{21}$$

where $\omega(R) := (2R)^2$ and $\kappa = 2$ for Euclidean ball cases, and $\omega(R) := \cosh^2 R + \sinh^2 R$ and $\kappa = \frac{1}{2}$ for hyperbolic ball cases. See Appendix G for the definition of ν, which depends on $\mathrm{P}_{\mathcal{X}}$ and $|\mathcal{V}|$.

Remark 11 (Comparison of Theorem 9 to Corollary 10)**.**

(a) Theorem 9 can apply to the most natural case $\psi(t) = t$, while Corollary 10 cannot since $\gamma(t) = \sqrt{t}$ or $\gamma(t) = \operatorname{acosh} t$ is not Lipschitz continuous.

(b) No LRC evaluation in (Suzuki et al., 2021b). Hence we cannot derive a faster bound than $O(\frac{1}{\sqrt{S}})$ in their direction, while we did as in (iii) of Theorem 1 thanks to the LRC evaluation by Theorem 9.

(c) The bound in Theorem 9 is polynomial in R even for hyperbolic space, better than Corollary 10, which is exponential in R. The comparison regarding the dependency on $|\mathcal{V}|$ is complicated. If we regard other variables as constants, Theorem 9, which is $O(|\mathcal{V}|)$, is always better than Corollary 10 owing to the second term in Corollary 10. However, if S is sufficiently large, then the second term vanishes. In that case, the discussion depends on ν, which again depends on $\mathrm{P}_{\mathcal{X}}$. See Appendix G for detailed discussion. In any case, the bound in Theorem 1 is much better in practical evaluations as the following example shows owing to the difference in the dependency on R.

(d) For Example 1 with $\psi(t) = t$, Corollary 10 gives $S \geq 7.30 \times 10^{72}$, a much larger data size than that by Theorem 1, as a sufficient condition for the hyperbolic method to outperform Euclidean method.

7. Discussion on proof and future work

As we explained in the Introduction section, our idea is to regard each hypothesis function as a function of the $|\mathrm{C}_{\mathcal{V}}|$ distance values, each of which corresponds to a couple of entities. Specifically, the proof of Theorem 9 evaluates $\mathrm{Rad}_{\mathrm{P},S}\big(h_{\ell,\mathcal{F}'_r}\big)$, where $\mathcal{F}'_r$ is given by replacing the condition $f \in \mathcal{F}$ in the definition of $\mathcal{F}_r$ by $\sum_{\{u,v\} \in \mathrm{C}_{\mathcal{V}}} \left(f(\{u,v\})\right)^2 \leq F^2$. Since $\mathcal{F}_r \subset \mathcal{F}'_r$, $\mathrm{Rad}_{\mathrm{P},S}\big(h_{\ell,\mathcal{F}'_r}\big) \leq \mathrm{Rad}_{\mathrm{P},S}\big(h_{\ell,\mathcal{F}'_r}\big)$ holds. Intuitively speaking, we allow any distance values that satisfy the condition about F, regardless of whether they are actually achievable by the representations in $\mathcal{W}$. This leads to an easy local Rademacher complexity evaluation. A potential issue here is that using $\mathcal{F}'_r$ might be too conservative since this function set has "forgot" the information that the hypothesis function comes from the representation space $\mathcal{W}$ and its distance function, other than it is restricted by F. Hence it is possible that $\mathcal{F}_r$ is no more than a very small part of $\mathcal{F}'_r$. If this is the case, we could improve our bound in the future.

Acknowledgements

Atsushi Suzuki is partially supported by Outstanding Foreign Scientist Studio (GZS2022012), Department of Science and Technology of Henan Province. Atsushi Nitanda is partially supported by JSPS Kakenhi (22H03650). Taiji Suzuki is partially supported by JST CREST. Kenji Yamanishi is partially supported by JSPS Kakenhi (19H01114).

References

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.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5: 135–146, 2017.

Bordes, A., Usunier, N., García-Durán, A., Weston, J., and Yakhnenko, O. Translating embeddings for modeling multi-relational data. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems*, pp. 2787–2795, 2013.

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.

Cortes, C. and Vapnik, V. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Dale, J. M., Popescu, L., and Karp, P. D. Machine learning methods for metabolic pathway prediction. *BMC bioinformatics*, 11(1):1–14, 2010.

Deza, M. M., Laurent, M., and Weismantel, R. *Geometry of cuts and metrics*, volume 2. Springer, 1997.

Ebisu, T. and Ichise, R. TorusE: Knowledge graph embedding on a Lie group. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18)*, pp. 1819–1826, 2018.

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.

Gao, Y., Zhang, C., Peng, J., and Parameswaran, A. The importance of norm regularization in linear graph embedding: Theoretical analysis and empirical demonstration. *arXiv preprint arXiv:1802.03560*, 2018.

Gromov, M. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Grover, A. and Leskovec, J. node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 855–864, 2016.

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 *7th International Conference on Learning Representations*, 2019.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098, 2002.

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.

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

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Koltchinskii, V. and Beznosova, O. Exponential convergence rates in classification. In *Proceedings of International Conference on Computational Learning Theory*, pp. 295–307, 2005.

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

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.

MA Basher, A. R. and Hallam, S. J. Leveraging heterogeneous network embedding for metabolic pathway prediction. *Bioinformatics*, 37(6):822–829, 2021.

Mendelson, S. Geometric parameters of kernel machines. In *Proceedings of International conference on computational learning theory*, pp. 29–43, 2002.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Conference on Neural Information Processing Systems*, pp. 3111–3119, 2013.

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.

Nickel, M., Tresp, V., and Kriegel, H. A three-way model for collective learning on multi-relational data. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 809–816, 2011.

Nickel, M., Rosasco, L., and Poggio, T. A. Holographic embeddings of knowledge graphs. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*, pp. 1955–1961, 2016.

Pennington, J., Socher, R., and Manning, C. D. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.

Perozzi, B., Al-Rfou, R., and Skiena, S. DeepWalk: online learning of social representations. In *Proceedings of The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 701–710, 2014.

Riedel, S., Yao, L., McCallum, A., and Marlin, B. M. Relation extraction with matrix factorization and universal schemas. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 74–84, 2013.

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.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Steinwart, I. and Christmann, A. *Support vector machines*. Springer Science & Business Media, 2008.

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.

Suzuki, A., Nitanda, A., Wang, J., Xu, L., Yamanishi, K., and Cavazza, M. Generalization error bound for hyperbolic ordinal embedding. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 10011–10021, 2021a.

Suzuki, A., Nitanda, A., Wang, J., Xu, L., Yamanishi, K., and Cavazza, M. Generalization bounds for graph embedding using negative sampling: Linear vs hyperbolic. In *Proceedings of the 35th Conference on Neural Information Processing Systems*, pp. 1243–1255, 2021b.

Tang, J., Qu, M., and Mei, Q. PTE: predictive text embedding through large-scale heterogeneous text networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1165–1174, 2015a.

Tang, J., Qu, M., Wang, M., Zhang, M., Yan, J., and Mei, Q. LINE: large-scale information network embedding. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015, Florence, Italy, May 18-22, 2015*, pp. 1067–1077, 2015b.

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

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. Complex embeddings for simple link prediction. In *Proceedings of the 33rd International Conference on Machine Learning,*, pp. 2071–2080, 2016.

Wang, Y., Wang, Y., Liu, X., and Pu, J. On the erm principle with networked data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Supplementary Materials
for Generalization Error Bound for Hyperbolic Ordinal Embedding

A. Proof of Theorem 1

We first confirm fundamental theorems to obtain excess risk bound from the Rademacher complexity.

Corollary 11 (Corollary from Theorem 3 in (Kakade et al., 2008))**.** *Suppose that the conditions (C1) and (C7) holds. Fix $\delta \in \mathbb{R}_{>0}$ and $\epsilon \in \mathbb{R}_{\geq 0}$. Then every ϵ-CERM $\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^S \to \mathbb{R}$ satisfies*

$$[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}\Big(\mathfrak{A}\Big(\left(z_s\right)_{s=1}^{S}\Big)\Big) - \inf\left\{\mathscr{R}_{\ell,\mathrm{P}}(f) | f \in \mathcal{F}\right\} \leq 2\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}}) + 2B\sqrt{\frac{\ln\frac{1}{\delta}}{S}} + \epsilon, \tag{22}$$

in probability at least $1 - \delta$.

The convergence rate of the bound given by the above corollary Corollary 11 is at the fastest $O\left(\frac{1}{\sqrt{S}}\right)$. Theorem 1 (ii) is derived using Corollary 11.

On the other hand, the other type of the excess risk bound, explained below, can give faster rate with some additional conditions. It uses the Rademacher complexity of a localized hypothesis function set, often called the ***local Rademacher complexity*** (Bartlett et al., 2005; Koltchinskii, 2006). The following is a simplified version of the version in (Steinwart & Christmann, 2008).

Corollary 12 (A simplified version of Theorem 7.20 in (Steinwart & Christmann, 2008))**.** *Let $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$ be equipped with a complete, separable metric dominating the pointwise convergence. Assume that conditions (C5) to (C9) and (9) are satisfied and fix L, B, V that satisfy the inequalities there. Also, assume that there exists a Bayes decision function $f^* \in \mathcal{L}_0(\mathcal{X})$, which satisfies $[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f^*) = [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M}$. Define the approximation error $\rho := \inf\left\{[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \middle| f \in \mathcal{F}\right\}$. For $r \geq \rho$, define $\mathcal{F}_r := \left\{f \in \mathcal{F} \middle| [\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \leq r\right\}$. Fix $f_0 \in \mathcal{F}$ and $B_0 > \sup\left\{\ell(x, y, f_0(x)) | (x, y) \in \mathcal{X} \times \mathcal{Y}\right\} \vee B$. Fix $S \in \mathbb{Z}_{>0}$, and assume that there exists a function $\varphi_S : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ that satisfies $\varphi_S(4r) \leq 2\varphi_S(r)$ and $\varphi_S(r) \geq \mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}_r})$. Fix $\delta \in \mathbb{R}_{>0}$, $\epsilon \in \mathbb{R}_{\geq 0}$, and $r \geq 30\varphi(r) \vee \left(\frac{72V \ln\frac{3}{\delta}}{S}\right)^{\frac{1}{2-\vartheta}} \vee \frac{5B_0 \ln\frac{3}{\delta}}{S} \vee \rho$. Then every ϵ-CERM $\mathfrak{A} : (\mathcal{X} \times \mathcal{Y})^S \to \mathbb{R}$ satisfies*

$$[\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}\Big(\mathfrak{A}\Big(\left(z_s\right)_{s=1}^{S}\Big)\Big) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \leq 6\big(\mathscr{R}_{\ell,\mathrm{P}}(f_0) - \mathscr{R}_{\ell,\mathrm{P}}^*\big) + 3r + 3\epsilon, \tag{23}$$

in probability at least $1 - \delta$.

Proof of Theorem 1. Since $\mathcal{X}$ is a finite sum, the expected risk is a finite weighted average of the loss. Since the loss function ℓ, the function ψ, and the distance function $\Delta_{\mathcal{W}}$ are all continuous from the assumption (C3), (C4) and (C8), we can regard the risk function $\mathscr{R}_{\ell,\mathrm{P}}$ is a continuous real function on $\mathcal{W}^{|\mathcal{V}|}$. Since $\mathcal{W}^{|\mathcal{V}|}$ is a compact topological space from the assumption (C2), the image of $\mathscr{R}_{\ell,\mathrm{P}}$ is also compact. Hence, we have a 0-CERM. Since $\mathcal{X}$ is a finite set, any map from $\mathcal{X}^S$ is measurable. In particular, the 0-CERM is measurable. It implies the statement (i) of Theorem 1.

The statement (ii) of Theorem 1 is the direct consequence of Corollary 11 if we admit Theorem 9, which we prove in the next section.

To prove the statement (iii) of Theorem 1, we need to show that $\mathcal{F} = \mathcal{F}_{\mathcal{W},\psi}$ is equipped with a complete, separable metric dominating the pointwise convergence. Since $\mathcal{X}$ is a finite set, we can regard $\mathcal{F}_{\mathcal{W},\psi} \subset \mathcal{L}_0(\mathcal{X})$ as a subset of $|\mathcal{X}|$-dimensional vector space. If we consider i.e., a standard Euclidean metric in the $|\mathcal{X}|$-dimensional vector space, it is obvious that is dominates the pointwise convergence and $\mathcal{F}_{\mathcal{W},\psi}$ is separable by the metric. Also, under the metric, the map $\mathrm{w} \mapsto f_{\mathrm{w},\psi}$ is continuous from the continuity of ψ and Δ_{w}. Here, we consider the topology of $\mathcal{W}^{\mathcal{V}}$ by identifying it with $\mathcal{W}^{|\mathcal{V}|}$. Since $\mathcal{F}_{\mathcal{W},\psi}$ is the image of the compact set $\mathcal{W}^{\mathcal{V}}$ by the above continuous map, $\mathcal{F}_{\mathcal{W},\psi}$ is also compact. This implies that $\mathcal{F}_{\mathcal{W},\psi}$ is complete (and totally bounded).

What remains to consider is the selection of φ_S. According to Theorem 9, we have that $\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}_r}) \leq \frac{\chi(r)}{\sqrt{S}}$,, where $\chi(r) := \min\{\chi_i(r) | i = 0, 1, \ldots, |\mathrm{C}_\mathcal{V}|\}$. However, if we substitute $\varphi_S(r)$ with $\frac{\chi(r)}{\sqrt{S}}$, it does not satisfy $\varphi_S(4r) \leq 2\varphi_S(r)$. Hence, we cannot directly apply Corollary 12 by substituting $\varphi_S(r)$ with $\frac{\chi(r)}{\sqrt{S}}$.

Nevertheless, we can see the following. There exist $a, b, d \in \mathbb{R}_{\geq 0}$ such that $\varphi_S(r) := a\big(r^\vartheta + d\big)^{\frac{1}{2\vartheta}} + b$ satisfies $\varphi_S(4r) \leq 2\varphi_S(r)$, $\varphi_S(4r) \geq \frac{\chi(r)}{\sqrt{S}}$, and $r \geq 30\varphi_S(r) \Leftrightarrow r \geq 30\frac{\chi(r)}{\sqrt{S}}$, for all $r \in \mathbb{R}_{\geq 0}$. Specifically, we can find such a, b, d as follows. First, recall that $r_i(S)$ for $1, \ldots |\mathrm{C}_\mathcal{V}|$ is defined as the only positive solution of the equation $r = {}^{30\chi_i(r)}/\sqrt{S}$ for r, where

$$\chi_i(r) := 2L\sqrt{2F^2\left(\frac{Ur^\vartheta}{4F^2}i + \mu_{\mathrm{P}_\mathcal{X}}(|\mathrm{C}_\mathcal{V}| - i)\right)} = c\sqrt{r^\vartheta + d}, \tag{24}$$

with $c = 2L\sqrt{2F^2\frac{U}{4F^2}i}$ and $d = \frac{4F^2}{U}\mu_{\mathrm{P}_\mathcal{X}}(|\mathrm{C}_\mathcal{V}| - i)$. Then, we can see that $(a, b, d) = \Big(c\vartheta(r_i(S))^{-\frac{1-\vartheta}{2}}, c(1-\vartheta)(r_i(S))^{\frac{\vartheta}{2}}, d\Big)$ satisfying the above properties, noting that $\varphi_S(r_i(S)) = \frac{\chi(r_i(S))}{\sqrt{S}}$ and $\frac{\mathrm{d}}{\mathrm{d}r}\varphi_S(r_i(S)) = \frac{\mathrm{d}}{\mathrm{d}r}\frac{\chi(r_i(S))}{\sqrt{S}}$ are satisfied with the (a, b, d).

Substituting such a φ_S and $f_0 = f^* \in \mathcal{F}$ in Corollary 12, we complete the proof of (iii) of Theorem 1. Note that $f^* \in \mathcal{F}$ is guaranteed by the condition (C9). □

B. Proof of Theorem 9

We review some basic properties of the Rademacher complexity.

Lemma 13. *Let $c \in \mathbb{R}$, $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$, and $f' \in \mathcal{L}_0(\mathcal{X})$. Then,*

$$\mathrm{Rad}_{\mathrm{P},S}(\{cf + f' | f \in \mathcal{F}\}) = |c|\mathrm{Rad}_{\mathrm{P},S}(\mathcal{F}). \tag{25}$$

For the proof of Lemma 13, see (e.g., Lemma 26.6, Shalev-Shwartz & Ben-David, 2014).

Lemma 14. *Let $\phi : \mathbb{R} \to \mathbb{R}$ be a Lipschitz continuous function and $\mathcal{F} \subset \mathcal{L}_0(\mathcal{X})$. Then, $\mathrm{Rad}_{\mathrm{P},S}(\{\phi \circ f | f \in \mathcal{F}\}) = \mathrm{Lip}\,(\phi)\mathrm{Rad}_{\mathrm{P},S}(\mathcal{F})$, where $\mathrm{Lip}\,(\phi) \in \mathbb{R}_{\geq 0}$ is the Lipschitz constant of ϕ.*

For the proof of Lemma 14, see (e.g., Lemma 26.9, Shalev-Shwartz & Ben-David, 2014). The following is easy using Lemma 14.

Lemma 15. *Suppose that the conditions (C6) and (C8) hold and L is a constant that satisfies the inequality in (C8).*

$$\mathrm{Rad}_{\mathrm{P},S}(h_{\ell,\mathcal{F}}) \leq L\mathrm{Rad}_{\mathrm{P}_\mathcal{X},S}(\mathcal{F}). \tag{26}$$

Proof of Theorem 9. We regard every element in $\mathcal{F}_{\mathcal{W},\psi}$ as a $|\mathrm{C}_\mathcal{V}|$-dimensional vector as follows. First, we fix an index map $\mathrm{ind} : \{1, 2, \ldots, |\mathrm{C}_\mathcal{V}|\} \to \mathrm{C}_\mathcal{V}$. We can use any map as ind as long as it is bijective. In the following, for a vector $\boldsymbol{u}$, we denote the i-th element by $[\boldsymbol{u}]_i$. We define $\boldsymbol{f}_{\mathrm{w},\psi} \in \mathbb{R}^{|\mathrm{C}_\mathcal{V}|}$ by $\left[\boldsymbol{f}_{\mathrm{w},\psi}\right]_i = f_{\mathrm{w},\psi}(\mathrm{ind}\,(i))$. Also, define $\boldsymbol{e}_{\{u,v\}} \in \mathbb{R}^{|\mathrm{C}_\mathcal{V}|}$ by

$$\left[\boldsymbol{e}_{\{u,v\}}\right]_i = \begin{cases} 1 & \text{if} \quad \mathrm{ind}\,(i) = \{u, v\}, \\ 0 & \text{if} \quad \mathrm{ind}\,(i) \neq \{u, v\}. \end{cases} \tag{27}$$

Since $\boldsymbol{f}_{\mathrm{w},\psi}^\top \boldsymbol{e}_{\{u,v\}} = f_{\mathrm{w},\psi}(\{u, v\})$, we can identify $\boldsymbol{f}_{\mathrm{w},\psi}$ and $\boldsymbol{x}_{\{u,v\}}$ with $f_{\mathrm{w},\psi}$ and $\{u, v\}$, respectively. For $f : \mathcal{X} \to \mathbb{R}$, we define $[f]_{-M}^{+M} : \mathcal{X} \to \mathbb{R}$ by $[f]_{-M}^{+M}(x) = [f(x)]_{-M}^{+M}$.

Recall

$$\mathcal{F}_r = \left\{ [f_{\mathrm{w},\psi}]_{-M}^{+M} \middle| [\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \leq r. \quad \right\}, \tag{28}$$

and

$$\mathcal{F}_r' = \left\{ [f]_{-M}^{+M} \middle| \begin{array}{r} f : \mathcal{X} \to \mathbb{R}, \\ \displaystyle\sum_{x \in \mathrm{C}_\mathcal{V}} (f(x))^2 \leq F^2, \\ [\mathscr{R}_{\ell,\mathrm{P}}]_{-M}^{+M}(f) - [\mathscr{R}_{\ell,\mathrm{P}}^*]_{-M}^{+M} \leq r. \end{array} \right\}. \tag{29}$$

Here, we have $\mathcal{F} \subset \mathcal{F}'$.

Define

$$\mathcal{F}_r^{(2)} := \left\{ [f]_{-M}^{+M} \;\middle|\; \begin{array}{r} f : \mathcal{X} \to \mathbb{R}, \\ \sum_{x \in \mathrm{C}_{\mathcal{V}}} (f(x))^2 \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[[f(x)]_{-M}^{+M} - [f^*(x)]_{-M}^{+M} \right]^2 \leq U r^{\vartheta}. \end{array} \right\}, \tag{30}$$

then $\mathcal{F}_r' \subset \mathcal{F}_r^{(2)}$ follows the condition (C10).

Using the vector notation, we have that

$$\begin{aligned} \mathcal{F}_r^{(2)} &= \left\{ {[\boldsymbol{f}]_{-M}^{+M}}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[[\boldsymbol{f}^{\top} \boldsymbol{e}_x]_{-M}^{+M} - [{\boldsymbol{f}^*}^{\top} \boldsymbol{e}_x]_{-M}^{+M} \right]^2 \leq U r^{\vartheta} \end{array} \right\} \\ &= \left\{ {[\boldsymbol{f}]_{-M}^{+M}}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[{[\boldsymbol{f}]_{-M}^{+M}}^{\top} \boldsymbol{e}_x - {[\boldsymbol{f}^*]_{-M}^{+M}}^{\top} \boldsymbol{e}_x \right]^2 \leq U r^{\vartheta} \end{array} \right\}, \end{aligned} \tag{31}$$

where we define $\boldsymbol{f}^* \in \mathbb{R}^{|\mathrm{C}_{\mathcal{V}}|}$ by $[\boldsymbol{f}^*]_i = f^*(\mathrm{ind}\,(i))$ and for $\boldsymbol{f} \in \mathbb{R}^{|\mathrm{C}_{\mathcal{V}}|}$ we define $[\boldsymbol{f}]_{-M}^{+M} \in \mathbb{R}^{|\mathrm{C}_{\mathcal{V}}|}$ by $\left[[\boldsymbol{f}]_{-M}^{+M} \right]_i = [[\boldsymbol{f}]_i]_{-M}^{+M}$.

Since $\boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2 \Rightarrow {[\boldsymbol{f}]_{-M}^{+M}}^{\top} [\boldsymbol{f}]_{-M}^{+M} \leq F^2$, we have that $\mathcal{F}_r^{(2)} \subset \mathcal{F}_r^{(3)}$, where $\mathcal{F}_r^{(3)}$ is defined by

$$\mathcal{F}_r^{(3)} := \left\{ \boldsymbol{f}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[\boldsymbol{f}^{\top} \boldsymbol{e}_x - {\boldsymbol{f}^*}^{\top} \boldsymbol{e}_x \right]^2 \leq U r^{\vartheta} \end{array} \right\}. \tag{32}$$

By Lemma 13, we have $\mathrm{Rad}_{\mathrm{P}_{\mathcal{X}}, S}\left(\mathcal{F}_r^{(3)}\right) = \mathrm{Rad}_{\mathrm{P}_{\mathcal{X}}, S}\left(\mathcal{F}_r^{(4)}\right)$, where $\mathcal{F}_r^{(4)}$ is given by

$$\mathcal{F}_r^{(4)} := \left\{ (\boldsymbol{f} - \boldsymbol{f}^*)^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[\boldsymbol{f}^{\top} \boldsymbol{e}_x - {\boldsymbol{f}^*}^{\top} \boldsymbol{e}_x \right]^2 \leq U r^{\vartheta} \end{array} \right\} \subset \tag{33}$$

We can evaluate the above set as follows.

$$\begin{aligned} \mathcal{F}_r^{(4)} &\subset \left\{ \left(\boldsymbol{f} - \boldsymbol{f}'\right)^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, {\boldsymbol{f}'}^{\top} \boldsymbol{f}' \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[\left(\boldsymbol{f} - \boldsymbol{f}'\right)^{\top} \boldsymbol{e}_x \right]^2 \leq U r^{\vartheta} \end{array} \right\} \\ &= \left\{ 2\boldsymbol{f}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq F^2, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[2\boldsymbol{f}^{\top} \boldsymbol{e}_x \right]^2 \leq U r^{\vartheta} \end{array} \right\} \\ &= \left\{ 2F\boldsymbol{f}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq 1, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[\boldsymbol{f}^{\top} \boldsymbol{e}_x \right]^2 \leq \dfrac{U r^{\vartheta}}{4F^2} \end{array} \right\}. \end{aligned} \tag{34}$$

Hence, by defining $\widehat{\mathcal{F}_r}$ as

$$\widehat{\mathcal{F}_r} := \left\{ \boldsymbol{f}^{\top} \boldsymbol{e}_{(\cdot)} \;\middle|\; \begin{array}{r} \boldsymbol{f}^{\top} \boldsymbol{f} \leq 1, \\ \mathbb{E}_{x \sim \mathrm{P}_{\mathcal{X}}} \left[\boldsymbol{f}^{\top} \boldsymbol{e}_x \right]^2 \leq \dfrac{U r^{\vartheta}}{4F^2} \end{array} \right\}, \tag{35}$$

we have that $\mathrm{Rad}_{\mathrm{P}_{\mathcal{X}}, S}\left(\mathcal{F}_r^{(4)}\right) \leq 2F \mathrm{Rad}_{\mathrm{P}_{\mathcal{X}}, S}\left(\widehat{\mathcal{F}_r}\right)$ from Lemma 14.

We apply Theorem 41 in (Mendelson, 2002). The following is the version in (Bartlett et al., 2005) given as the first half of Theorem 6.5.

Theorem 16 (The first half of Theorem 6.5 in (Bartlett et al., 2005), given in Theorem 41 in (Mendelson, 2002)for the first time.)**.** *Let $\mathcal{X}$ be a measurable set and P is a distribution on it. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a positive semidefinite kernel function that satisfies $\mathbb{E}_{x\sim\mathrm{P}} k(x,x) < +\infty$. Define the integral operator $T : \mathcal{L}_2(\mathrm{P}) \to \mathcal{L}_2(\mathrm{P})$ by $(T(f))(x) \coloneqq \mathbb{E}_{x'\sim\mathrm{P}} k(x,x')f(x')$ and let $(\lambda_i)_{i=1}^{\infty}$ be the sequence of the eigenvalues of T. Let $\mathscr{H}_k$ be the reproducing kernel Hilbert space generated by k and denote its norm function by $\|\cdot\|_{\mathscr{H}_k}$. Then,*

$$\mathrm{Rad}_{\mathrm{P},S}\left(\left\{f \in \mathscr{H}_k \,\middle|\, \|f\|_{\mathscr{H}_k} \le 1, \mathbb{E}_{x\sim\mathrm{P}}(f(x))^2 \le \rho\right\}\right) \le \sqrt{\frac{2}{S}\sum_{i=1}^{\infty} \min\{\rho, \lambda_i\}}. \tag{36}$$

Here, we consider the linear kernel function $k(\boldsymbol{x},\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{x}$. Then we can easily confirm $\left\|\boldsymbol{f}^\top\boldsymbol{e}_{(\cdot)}\right\|_{\mathscr{H}_k} = \sqrt{\boldsymbol{f}^\top\boldsymbol{f}}$, and T is given by the matrix $\sum_{i=1}^{|\mathrm{C}_{\mathcal{V}}|} \mathrm{P}_{\mathcal{X}}(\{\mathrm{ind}\,(i)\})\boldsymbol{e}_{(\mathrm{ind}\,(i))}\boldsymbol{e}_{(\mathrm{ind}\,(i))}^\top$. Hence, we have that

$$\lambda_i = \begin{cases} \mathrm{P}_{\mathcal{X}}(\{\mathrm{ind}\,(i)\}) & \text{if} \quad i = 1, 2, \dots, |\mathrm{C}_{\mathcal{V}}|, \\ 0 & \text{if} \quad i > |\mathrm{C}_{\mathcal{V}}|. \end{cases} \tag{37}$$

Applying the above and using Lemma 15, we complete the proof. □

C. Proof of Lemma 2

Proof of Lemma 2. We first prove it for $\tau = 1$. Let $\boldsymbol{w}_v \in \mathbb{R}^D$ be the representation of $v \in \mathcal{V}$. Fix a bijective map $\mathrm{ind} : \{1, 2, \dots, |\mathrm{C}_{\mathcal{V}}|\} \to \mathrm{C}_{\mathcal{V}}$, which we call an indexing map. We define the representation matrix $\boldsymbol{W} \in \mathbb{R}^{D,|\mathcal{V}|}$ by $\boldsymbol{W} \coloneqq \begin{bmatrix}\boldsymbol{w}_{\mathrm{ind}\,(1)} & \boldsymbol{w}_{\mathrm{ind}\,(2)} & \cdots & \boldsymbol{w}_{\mathrm{ind}\,(|\mathcal{V}|)}\end{bmatrix}$. Then,

$$\begin{aligned} &\sum_{\{u,v\}\in|\mathrm{C}_{\mathcal{V}}|} (\Delta_{\mathcal{W}}(\mathrm{w}(u), \mathrm{w}(v)))^2 \\ &= \frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{W}^\top\boldsymbol{W}\left(|\mathcal{V}|\mathbf{I}_{|\mathcal{V}|} - \mathbf{1}_{|\mathcal{V}|}\mathbf{1}_{|\mathcal{V}|}^\top\right)\right\} \\ &\le \frac{1}{2}\operatorname{Tr}\left\{\boldsymbol{W}^\top\boldsymbol{W}\left(|\mathcal{V}|\mathbf{I}_{|\mathcal{V}|}\right)\right\} \\ &= \frac{|\mathcal{V}|}{2}\operatorname{Tr}\left\{\boldsymbol{W}^\top\boldsymbol{W}\right\} \\ &= \frac{|\mathcal{V}|}{2}\sum_{v\in\mathcal{V}} \boldsymbol{w}_v^\top\boldsymbol{w}_v, \\ &\le \frac{|\mathcal{V}|}{2}R^2. \end{aligned} \tag{38}$$

If $\tau > 1$, it follows that

$$\begin{aligned} &\sum_{\{u,v\}\in|\mathrm{C}_{\mathcal{V}}|} (\Delta_{\mathcal{W}}(\mathrm{w}(u), \mathrm{w}(v)))^{2\tau} \\ &\le \sum_{\{u,v\}\in|\mathrm{C}_{\mathcal{V}}|} (\Delta_{\mathcal{W}}(\mathrm{w}(u), \mathrm{w}(v)))^2 (2R)^{2(\tau-1)}, \end{aligned} \tag{39}$$

which completes the proof. □

D. Proof of Lemma 3

Proof of Lemma 3. First, we prove it for $\tau = 1$. Since $D \ge 2$, the space contains a two dimensional hyperbolic disk as a subspace. In the hyperbolic disk, consider a regular polygon centered at the origin with $|\mathcal{V}|$ vertices and radius R. Using the hyperbolic law of sines, we have that the length of one side in the polygon is given by $2\operatorname{asinh}\left(\sin\frac{\pi}{|\mathcal{V}|}\sinh R\right)$. Since $\frac{2\operatorname{asinh}\left(\sin\frac{\pi}{|\mathcal{V}|}\sinh R\right)}{R} \to 2$ as $R \to \infty$, we obtain the consequence of the lemma for $\tau = 1$. For $\tau > 1$, we obtain the consequence by (39), which completes the proof. □

E. Hinge loss and Corollary 4

In this section, we just confirm that Corollary 4 immediately follows Theorem 1 and the following existing theorem.

Theorem 17 (Theorem 8.24 in (Steinwart & Christmann, 2008))**.** *Let* P *be a distribution on* $\mathcal{X} \times \{\pm 1\}$ *and the loss function be the hinge loss* $\ell_{\mathrm{hinge}}(x, y, t) := \phi_{\mathrm{hinge}}(yt)$*, where* $\phi_{\mathrm{hinge}}(t') := \max\{1 - t', 0\}$ *with* $M = 1$*. Define the risk function* $\mathscr{R}_{\ell,\mathrm{P}}$ *as in Section 2.5. Assume that the distribution* P *has noise exponent* $q \in \mathbb{R}_{\geq 0}$ *with constant* $c \in \mathbb{R}_{>0}$*. Then, for all* $f \in \mathcal{L}_0(\mathcal{X})$*, then the condition (C10) in Assumption 1 holds with* $\vartheta = \frac{q}{q+1}$ *and* $U = 6c^{\frac{q}{q+1}}$*.*

F. General condition for hyperbolic to outperform Euclidean

In Example 1, we gave the condition for hyperbolic graph embedding to outperform Euclidean graph embedding on a specific setting. We give the condition for a general setting in the following, which we can obtain by simple calculation from Theorem 1.

Proposition 18. *Suppose that conditions (C1) to (C4) in Assumption 1 are satisfied, the loss function be the hinge loss, and* $\psi(t) = t^{\tau}$*. Let the true dissimilarity* $\Delta^{*} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}_{\geq 0}$ *be given by the graph distance of a tree. Then, for* R *given by Lemma 5, the expected risk of a CERM using* $\mathcal{B}[R; \mathcal{H}^2]$ *is better than any CERM using* $\mathcal{R}^2$ *in probability at least* $1 - \delta$ *if* $S \geq \left(r_0 \wedge r_{|\mathrm{C}_{\mathcal{V}}|}\right) \vee a$*, where*

$$\begin{aligned}
r_0 &:= 97200\left[\tau(2R)^{\tau-1}\right]^2 |\mathrm{C}_{\mathcal{V}}|^2 \frac{1}{\xi^2 v_{\mathit{min}}(R; \mathcal{R}^2)},\\
r_{|\mathrm{C}_{\mathcal{V}}|} &:= 32R^2\left[\tau(2R)^{\tau-1}\right]^2 |\mathrm{C}_{\mathcal{V}}|^2 \frac{1}{\xi^2 [v_{\mathit{min}}(R; \mathcal{R}^2)]^2},\\
a &:= 3888\frac{1}{\xi^2 v_{\mathit{min}}(R; \mathcal{R}^2)} \ln\frac{3}{\delta}.
\end{aligned} \tag{40}$$

G. The definition of ν and dependency of the bounds by Theorem 1 and Corollary 10 on $|\mathcal{V}|$.

The value ν, which the bound in Corollary 10 depends on, is defined in (Suzuki et al., 2021b) as $\nu := \left\|\mathbb{E}_{\{u,v\}\sim\mathrm{P}_{\mathcal{X}}} \boldsymbol{E}_{\{u,v\}}^2\right\|_{\mathrm{op},2}$, where the symmetric matrix $\boldsymbol{E}_{\{u,v\}}$ for $\{u, v\} \in \mathrm{C}_{\mathcal{V}}$ is given by

$$\left[\boldsymbol{E}_{\{u,v\}}\right]_{i,j} = \begin{cases} c_{\mathrm{diag}} & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} \subsetneq \{u, v\}, \\ c_{\mathrm{off}} & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} = \{u, v\}, \\ 0 & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} \not\subset \{u, v\}. \end{cases} \tag{41}$$

Here $(c_{\mathrm{diag}}, c_{\mathrm{off}}) = (1, -1)$ for the Euclidean case, and $(c_{\mathrm{diag}}, c_{\mathrm{off}}) = \left(0, -\frac{1}{2}\right)$. Here, $\|\cdot\|_{\mathrm{op},2}$ is the operator norm with respect to 2-norm. For a real symmetric matrix $\boldsymbol{A}$, $\|\boldsymbol{A}\|_{\mathrm{op},2}$ equals to the maximum eigenvalue of $\boldsymbol{A}$ and also equals to the maximum singular value of $\boldsymbol{A}$. We have that

$$\left[\boldsymbol{E}_{\{u,v\}}^2\right]_{i,j} = \begin{cases} c'_{\mathrm{diag}} & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} \subsetneq \{u, v\}, \\ c'_{\mathrm{off}} & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} = \{u, v\}, \\ 0 & \text{if} \quad \{\mathrm{ind}\,(i), \mathrm{ind}\,(j)\} \not\subset \{u, v\}, \end{cases} \tag{42}$$

where $(c_{\mathrm{diag}}, c_{\mathrm{off}}) = (2, -2)$ for the Euclidean case, and $(c_{\mathrm{diag}}, c_{\mathrm{off}}) = \left(\frac{1}{4}, 0\right)$. For the upper bound of ν, as pointed out by (Suzuki et al., 2021b), we have that $\nu := \left\|\mathbb{E}_{\{u,v\}\sim\mathrm{P}_{\mathcal{X}}} \boldsymbol{E}_{\{u,v\}}^2\right\|_{\mathrm{op},2} \leq \mathbb{E}_{\{u,v\}\sim\mathrm{P}_{\mathcal{X}}} \left\|\boldsymbol{E}_{\{u,v\}}^2\right\|_{\mathrm{op},2}$ from Jensen's inequality. The right side is 4 for the Euclidean case and $\frac{1}{4}$ for the hyperbolic case. Indeed, these upper bounds are achievable if only one couple of entities is generated. For the lower bound, we can see that the trace of $\boldsymbol{E}_{\{u,v\}}^2$ is always 4 for the Euclidean case and $\frac{1}{2}$ for hyperbolic case, as we can see by summing the diagonal elements up. Hence, it also holds for its expectation $\mathbb{E}_{\{u,v\}\sim \mathrm{P}_{\mathcal{X}}} \boldsymbol{E}^2_{\{u,v\}}$. We remark that the trace equals to the sum of eigenvalues. Since we have $|\mathrm{C}_{\mathcal{V}}|$ eigenvalues, the mean of eigenvalues is $\frac{4}{|\mathrm{C}_{\mathcal{V}}|}$ for the Euclidean case and $\frac{1}{2|\mathrm{C}_{\mathcal{V}}|}$ for the hyperbolic case. The value ν is the maximum in the eigenvalues, which is not smaller than the mean. Hence, ν is lower-bounded by $\frac{4}{|\mathrm{C}_{\mathcal{V}}|}$ for the Euclidean case and $\frac{1}{2|\mathrm{C}_{\mathcal{V}}|}$ for the hyperbolic case. For both cases, the lower-bound is achieved by the uniform distribution.

Let us consider the bound by Corollary 10 again. If we focus on $|\mathcal{V}|$ and S, the bound is $O\left(\frac{|\mathcal{V}|\sqrt{\nu \ln |\mathcal{V}|}}{\sqrt{S}} + \frac{\kappa |\mathcal{V}| \ln |\mathcal{V}|}{S}\right)$. For the upper bound case, Corollary 10 gives $O\left(\frac{|\mathcal{V}|\sqrt{\ln |\mathcal{V}|}}{\sqrt{S}} + \frac{\kappa |\mathcal{V}| \ln |\mathcal{V}|}{3S}\right)$. Since Theorem 9 gives the bound that is $O\left(\frac{\sqrt{|\mathcal{V}|}}{\sqrt{S}}\right)$ for the Euclidean case and $O\left(\frac{|\mathcal{V}|}{\sqrt{S}}\right)$ for the hyperbolic case, Theorem 9 is better than Corollary 10. For the lower bound case, Corollary 10 gives $O\left(\frac{\sqrt{\ln |\mathcal{V}|}}{\sqrt{S}} + \frac{\kappa |\mathcal{V}| \ln |\mathcal{V}|}{3S}\right)$. Here, the dependency on $|\mathcal{V}|$ is significantly different between the first and second term. It implies that if S is sufficiently large, then Corollary 10 is better in the dependency on $|\mathcal{V}|$ than Theorem 9, while the converse holds if S is not large.