Lifting Weak Supervision To Structured Prediction

Anonymous Author(s)
Affiliation
Address
email

Abstract

Weak supervision (WS) is a rich set of techniques that produce pseudolabels by aggregating easily obtained but potentially noisy label estimates from various sources. WS is theoretically well-understood for binary classification, where simple approaches enable consistent estimation of pseudolabel noise rates. Using this result, it has been shown that downstream models trained on the pseudolabels have generalization guarantees nearly identical to those trained on clean labels. While this is exciting, users often wish to use WS for *structured prediction*, where the output space consists of more than a binary or multi-class label set: e.g. rankings, graphs, manifolds, and more. Do the favorable theoretical properties of WS for binary classification lift to this setting? We answer this question in the affirmative for a wide range of scenarios. For labels taking values in a finite metric space, we introduce techniques new to weak supervision based on pseudo-Euclidean embeddings and tensor decompositions, providing a nearly-consistent noise rate estimator. For labels in constant-curvature Riemannian manifolds, we introduce new invariants that also yield consistent noise rate estimation. In both cases, when using the resulting pseudolabels in concert with a flexible downstream model, we obtain generalization guarantees nearly identical to those for models trained on clean data. Several of our results, which can be viewed as robustness guarantees in structured prediction with noisy labels, may be of independent interest.

1 Introduction

Weak supervision (WS) is an array of methods used to construct pseudolabels for training supervised models in label-constrained settings. The standard workflow [RSW$^+$16, RBE$^+$18, FCS$^+$20] is to assemble a set of cheaply-acquired labeling functions—simple heuristics, small programs, pretrained models, knowledge base lookups—that produce multiple noisy estimates of what the true label is for each unlabeled point in a training set. These noisy outputs are modeled and aggregated into a single higher-quality pseudolabel. Any conventional supervised end model can be trained on these pseudolabels. This pattern has been used to deliver excellent performance in a range of domains in both research and industry settings [DRS$^+$20, RNGS20, SLB20], bypassing the need to invest in large-scale manual labeling. Importantly, these successes are usually found in binary or small-cardinality classification settings.

While exciting, users often wish to use weak supervision in *structured prediction* (SP) settings, where the output space consists of more than a binary or multiclass label set [BHS$^+$07, KL15]. In such cases, there exists meaningful algebraic or geometric structure to exploit. Structured prediction includes, for example, learning rankings used for recommendation systems [KAG18], regression in metric spaces [PM19], learning on manifolds [RCMR18], graph-based learning [GS19], and more.

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

An important advantage of WS in the standard setting of binary classification is that it yields models with nearly the same generalization guarantees as their fully-supervised counterparts. Indeed, the penalty for using pseudolabels instead of clean labels is only a multiplicative constant. This is a highly favorable tradeoff since acquiring more unlabeled data is easy. This property leads us to ask the key question for this work: **does weak supervision for structured prediction preserve generalization guarantees?** We answer this question in the affirmative, justifying the application of WS to settings far from its current use.

Generalization results in WS rely on two steps [RHD$^+$19, FCS$^+$20]: (i) showing that the estimator used to learn the model of the labeling functions is consistent, thus recovering the noise rates for these noisy voters, and (ii) using a noise-aware loss to de-bias end-model training [NDRT13]. Lifting these two results to structured prediction is challenging. The only available weak supervision technique suitable for SP is that of [SLV$^+$22]. It suffers from several limitations. First, it relies on the availability of isometric embeddings of metric spaces into $\mathbb{R}^d$—but does not explain how to find these. Second, it does not tackle downstream generalization at all. We resolve these two challenges.

We introduce results for a wide variety of structured prediction problems, requiring only that the labels live in some metric space. We consider both finite and continuous (manifold-valued) settings. For finite spaces, we apply two tools that are new to weak supervision. The approach we propose combines isometric *pseudo-Euclidean embeddings* with *tensor decompositions*—resulting in a nearly-consistent noise rate estimator. In the continuous case, we introduce a label model suitable for the so-called *model spaces*—Riemannian manifolds of constant curvature—along with extensions to even more general spaces. In both cases, we show generalization results when using the resulting pseudolabels in concert with a flexible end model from [CRR16, RCMR18].

Contributions:

- New techniques for performing weak supervision in finite metric spaces based on isometric pseudo-Euclidean embeddings and tensor decomposition algorithms,
- Generalizations of weak supervision for regression to manifold-valued regression in constant-curvature manifolds,
- Finite-sample error bounds for noise rate estimation in each scenario,
- Generalization error guarantees for training downstream models on pseudolabels.

2 Background and Problem Setup

Our goal is to theoretically characterize how well we can learn with pseudolabels (built with weak supervision techniques) in structured prediction settings. Specifically, we seek to understand the interplay between the noise in WS sources and the generalization performance of the downstream structured prediction model. We provide a brief background on structured prediction and weak supervision, then introduce our problem and some useful notation.

2.1 Structured Prediction

Structured prediction (SP) involves predicting labels in spaces with rich structure. Denote the label space by $\mathcal{Y}$. Conventionally $\mathcal{Y}$ is a set, e.g., $\mathcal{Y} = \{-1, +1\}$ for binary classification. In the SP setting, $\mathcal{Y}$ has some additional algebraic or geometric structure. In this work we assume that $\mathcal{Y}$ is a metric space with metric (distance) $d_{\mathcal{Y}}$. This covers many types of problems, including

- Rankings, where $\mathcal{Y} = S_\rho$, the symmetric group on $\{1, \ldots, \rho\}$, i.e. labels are permutations,
- Graphs, where $\mathcal{Y} = \mathcal{G}_\rho$, the space of graphs with vertex set $V = \{1, \ldots, \rho\}$,
- Riemannian manifolds, where $\mathcal{Y} = \mathbb{S}_d$, the sphere, or $\mathbb{H}_d$, the hyperboloid.

In conventional supervised learning we have a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ of i.i.d samples drawn from some distribution ρ over the space $\mathcal{X} \times \mathcal{Y}$.

Learning and Generalization in SP As usual, we seek to learn a model that generalizes well to points not seen during training. Let $\mathcal{F} = \{f : \mathcal{X} \mapsto \mathcal{Y}\}$ be a family of functions from $\mathcal{X}$ to $\mathcal{Y}$. Define the risk $R(f)$ for $f \in \mathcal{F}$ and f^* as

$$R(f) = \int_{\mathcal{X}\times\mathcal{Y}} d^2_{\mathcal{Y}}(f(x), y) d\rho(x, y) \qquad f^* \in \arg\min_{f\in\mathcal{F}} R(f). \tag{1}$$

For a large class of settings (including all of those we consider in this paper), [CRR16, RCMR18] have shown that the estimator $\hat{f}$ defined below approaches f^*:

$$\hat{f}(x) = \arg\min_{y\in\mathcal{Y}} F(x, y) \qquad F(x, y) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x) d^2_{\mathcal{Y}}(y, y_i), \tag{2}$$

where $\alpha(x) = (\mathbf{K} + \nu\mathbf{I})^{-1}\mathbf{K}_x$. Here, $\mathbf{K}$ is the kernel matrix for a p.d. kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, so that $\mathbf{K}_{i,j} = k(x_i, x_j)$, $(\mathbf{K}_x)_i = k(x, x_i)$, and ν is a regularization parameter. Thus it is necessary to first learn the weights α and then to perform the optimization in (2) to make a prediction.

When there is no label noise, an exciting contribution of [CRR16, RCMR18] is the generalization bound

$$R(\hat{f}) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}),$$

that holds with high probability. The key question we tackle is *does the use of pseudolabels instead of true labels y_i affect the generalization rate?* Note that even having access to the kernel and thus knowing the weights α is insufficient to ensure this; the presence of noise when replacing y_i with a pseudolabel could ostensibly ruin the generalization bound.

2.2 Weak Supervision

In WS, we cannot access *any* of the ground-truth labels y_i. Instead we observe for each x_i the noisy votes $\lambda_{a,i}, \ldots, \lambda_{m,i}$. These are m weak supervision outputs provided by *labeling functions* (LFs) s_a, where $s_a : \mathcal{X} \to \mathcal{Y}$ and $\lambda_{a,i} = s_a(x_i)$. A two step process is used to construct pseudolabels. First, we learn a *noise model* (also called a label model) that determines how reliable each source s_a is. That is, we must learn $\boldsymbol{\theta}$ for $P_{\boldsymbol{\theta}}(\lambda_1, \lambda_2, \ldots, \lambda_m|y)$—without having access to any samples of y. Second, the noise model is used to infer a distribution (or its mode) for each point: $P_{\boldsymbol{\theta}}(y_i|\lambda_{1,i}, \ldots, \lambda_{m,i})$.

We adopt the noise model from [SLV$^+$22], which is suitable for our SP setting:

$$P_{\boldsymbol{\theta}}(\lambda_1, \ldots, \lambda_m|Y = y) = \frac{1}{Z}\exp\left(-\sum_{a=1}^{m} \theta_a d^2_{\mathcal{Y}}(\lambda_a, y)\right). \tag{3}$$

This is an exponential family model, where Z is the normalizing partition function and $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^T > 0$ are the *canonical* parameters. The model can also be described in terms of the *mean* parameters $\mathbb{E}[d^2_{\mathcal{Y}}(\lambda_a, y)]$. Intuitively, if θ_a is large, then the typical distance from λ_a to y must be small.In this case the LF is reliable. Conversely, if θ_a is small, the LF is unreliable.

Our goal is to form estimates $\hat{\boldsymbol{\theta}}$ in order to construct pseudolabels. One way to build such pseudolabels is to compute $\tilde{y} = \arg\min_{z\in\mathcal{Y}} 1/m \sum_{a=1}^{m} \hat{\theta}_a d^2_{\mathcal{Y}}(z, \lambda_a)$. Observe how the estimated parameters are used to weight the labeling functions θ_a, ensuring that more reliable votes receive a larger weight. The extreme cases are $\theta_a = 0$, so that λ_a is independent of y, and so gets no weight, and $\theta_a = \infty$, so that $\theta_a = y$ and should get all of the weight.

We are now in a position to state the main research question for this work:

Do there exist estimation approaches yielding $\hat{\theta}$ that produce pseudolabels $\tilde{y}$ that maintain the same generalization error rate $\mathcal{O}(n^{-1/4})$ when used in (2), or a modified version of (2)?

3 Noise Rate Recovery in Finite Metric Spaces

In the next two sections we will handle finite metric spaces. Afterwards we tackle continuous (manifold-valued) spaces. We first discuss learning the noise parameters $\boldsymbol{\theta}$, then the use of pseudolabels in training.

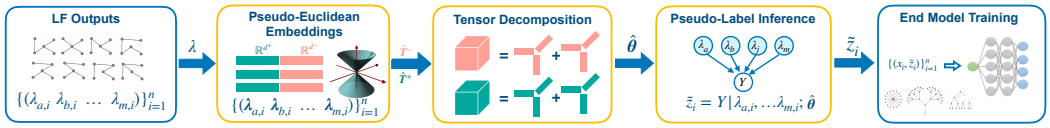

Figure 1: Illustration of our weak supervision pipeline for the finite label space setting.

Roadmap For finite metric spaces with $|\mathcal{Y}| = r$, we apply two tools new to weak supervision. First, we embed $\mathcal{Y}$ into a *pseudo-Euclidean* space [Gol85]. These spaces generalize Euclidean space, enabling isometric (distance-preserving) embeddings for any metric space. Using pseudo-Euclidean spaces make our analysis slightly more complex, but we gain the isometry property, which is critical.

Second, we form three-way tensors from embeddings of observed labeling functions. Applying tensor product decomposition algorithms [AGH+14], we can recover estimates of the mean parameters $\hat{\mathbb{E}}[d^2_{\mathcal{Y}}(\lambda_a, y)]$ and ultimately $\hat{\theta}_a$. Finally, we reweight the model (2) to preserve generalization.

3.1 Pseudo-Euclidean Embeddings

Working directly with the label space $\mathcal{Y}$ is challenging due to its potentially large cardinality. A standard way to address this challenge is to embed $\mathcal{Y}$ into a vector space. For example, multi-dimensional scaling (MDS) [KW78] embeds $\mathcal{Y}$ into $\mathbb{R}^d$. The downside of MDS is that only some metric spaces embed (isometrically) into Euclidean space. In particular, it is necessary that the square distance matrix $\mathbf{D}$ is positive semi-definite.

A simple and elegant way to overcome this difficulty is to instead use *pseudo-Euclidean* spaces for embeddings. These pseudo-spaces do not require a p.s.d. inner product. As an outcome, any finite metric space can be embedded into a pseudo-Euclidean space with *no distortion* [Gol85]—so that distances are exactly preserved. We shall need only a few properties of these spaces: A vector $\mathbf{u}$ in a pseudo-Euclidean space $\mathbb{R}^{d^+,d^-}$ has two parts $\mathbf{u}^+ \in \mathbb{R}^{d^+}$ and $\mathbf{u}^- \in \mathbb{R}^{d^-}$. The dot product and the squared distance between any two vectors $\mathbf{u}, \mathbf{v}$ in a pseudo-Euclidean space $\mathbb{R}^{d^+,d^-}$ are $\langle \mathbf{u}, \mathbf{v} \rangle_\phi = \langle \mathbf{u}^+, \mathbf{v}^+ \rangle - \langle \mathbf{u}^-, \mathbf{v}^- \rangle$ and $d^2_\phi(\mathbf{u}, \mathbf{v}) = ||\mathbf{u}^+ - \mathbf{v}^+||_2^2 - ||\mathbf{u}^- - \mathbf{v}^-||_2^2$.

Example: To see why such embeddings are advantageous, we compare using a one-hot vector representation (whose dimension is $|\mathcal{Y}|$) versus an embedding. Consider a tree with a root node and three branches, each of which is a path with t nodes. Let $\mathcal{Y}$ be the nodes in the tree with the shortest-hops distance as the metric. It can be shown (using [BS16]) that the pseudo-Euclidean embedding dimension is just $d = 3$. The one-hot embedding dimension is $d = |\mathcal{Y}| = 3t + 1$—arbitrarily larger!

Now we are ready to apply these embeddings to our problem. Abusing notation, we write $\boldsymbol{\lambda}_a$ and $\mathbf{y}$ for the pseudo-Euclidean embeddings of λ_a, y. We have that $d^2_{\mathcal{Y}}(\lambda_a, y) = d^2_\phi(\boldsymbol{\lambda}_a, \mathbf{y})$, so that there is no loss of information from working with these spaces. In addition, we write the mean as $\boldsymbol{\mu}_{a,y} = \mathbb{E}[\boldsymbol{\lambda}_a | \mathbf{y}]$ and the covariance as $\boldsymbol{\Sigma}_{a,y}$. It is easy to see that $\boldsymbol{\mu}_{a,y}$ can be used to obtain the mean parameters $\mathbb{E}[d^2_{\mathcal{Y}}(\lambda_a, y)]$ — due to the isometric distances and nice form of distance function in the pseudo-Euclidean spaces we can use $\boldsymbol{\mu}_{a,y}$ to bound $\mathbb{E}[d^2_\phi(\boldsymbol{\lambda}_a, \mathbf{y})]$. Thus our goal is to get an accurate estimate $\hat{\boldsymbol{\mu}}_{a,y} = \hat{\mathbb{E}}[\boldsymbol{\lambda}_a | \mathbf{y}]$. If we could observe y, it would be easy to form an empirical estimate of $\hat{\boldsymbol{\mu}}_{a,y}$—but we do not have access to it. Our approach will be to apply tensor decomposition approaches for multi-view mixtures.

3.2 Multi-View Mixtures and Tensor Decompositions

In a multi-view mixture model, multiple views $\{\lambda_a\}_{a=1}^m$ of a latent variable Y are observed. These views are independent when conditioned on Y. We treat the positive and negative components $\boldsymbol{\lambda}_a^+ \in \mathbb{R}^{d^+}$ and $\boldsymbol{\lambda}_a^- \in \mathbb{R}^{d^-}$ of our pseudo-Euclidean embedding as separate multi-view mixtures:

$$\boldsymbol{\lambda}_a^+ | \mathbf{y} \sim \boldsymbol{\mu}_{a,y}^+ + \sigma\sqrt{d^+} \cdot \boldsymbol{\epsilon}_a^+ \quad \text{and} \quad \boldsymbol{\lambda}_a^- | \mathbf{y} \sim \boldsymbol{\mu}_{a,y}^- + \sigma\sqrt{d^-} \cdot \boldsymbol{\epsilon}_a^- \qquad \forall a \in [m]. \tag{4}$$

where $\boldsymbol{\mu}_{a,y}^+ = \mathbb{E}[\boldsymbol{\lambda}_a^+ | \mathbf{y}]$, $\boldsymbol{\mu}_{a,y}^- = \mathbb{E}[\boldsymbol{\lambda}_a^- | \mathbf{y}]$ and $\boldsymbol{\epsilon}_a^+, \boldsymbol{\epsilon}_a^-$ are mean zero random vectors with covariances $\frac{1}{d^+}\mathbf{I}_{d^+}, \frac{1}{d^-}\mathbf{I}_{d^-}$ respectively. Here σ^2 is a proxy variance for the noise model in (3).

Algorithm 1 Algorithm for Pseudolabel Construction

Input: Labeling function outputs $\mathbf{L} = \{(\lambda_{a,i}, \lambda_{m,i})\}_{i=1}^n$, Label Space $\mathcal{Y} = \{y_0, \dots y_{r-1}\}$
Output: Pseudolabels for each data point $\mathbf{Z} = \{\tilde{z}_i\}_{i=1}^n$

▷ Step 1: Compute pseudo-Euclidean Embeddings
Compute pairwise distance matrix $\mathbf{D} \in \mathbb{R}^{r\times r}$ with $\mathbf{D}_{ij} = d^2_{\mathcal{Y}}(y_i, y_j)$
Construct matrix $\mathbf{M} \in \mathbb{R}^{r\times r}$ with $\mathbf{M}_{ij} = \frac{1}{2}(\mathbf{D}^2_{0i} + \mathbf{D}^2_{0j} - \mathbf{D}^2_{ij})$
Compute eigendecomposition of $\mathbf{M}$ and let $\mathbf{M} = \mathbf{U}\mathbf{C}\mathbf{U}^T$
Let l^+, l^- be the list of indices of positive and negative eigenvalues sorted by their magnitude.
Let $d^+ = |l^+|, \quad d^- = |l^-|$
Construct permutation matrix $\mathbf{I}_{perm} \in \mathbb{R}^{r\times(d^+ + d^-)}$ by concatenating l^+, l^- in order
$\bar{\mathbf{C}} = \mathbf{C}\mathbf{I}_{perm}, \bar{\mathbf{U}} = \mathbf{U}\mathbf{I}_{perm}$
$\mathbb{Y} = \bar{\mathbf{U}}^T \bar{\mathbf{C}}^{\frac{1}{2}} \in \mathbb{R}^{r\times(d^+ + d^-)}$ and let this define the mapping $g : \mathcal{Y} \mapsto \mathbb{Y}$

▷ Step 2: Parameter Estimation Using Tensor Decomposition
for $a \leftarrow 1$ to $m - 3$ **do**
    Obtain embeddings $\boldsymbol{\lambda}_{a,i} = g(\lambda_{a,i}), \boldsymbol{\lambda}_{a+1,i} = g(\lambda_{a+1,i}), \boldsymbol{\lambda}_{a+2,i} = g(\lambda_{a+2,i}) \quad \forall i \in [n]$
    Construct tensors $\hat{\mathbf{T}}^+$ and $\hat{\mathbf{T}}^-$ as defined in (5) for triple $(a, a+1, a+2)$
    $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^+_{a+1,y}, \hat{\boldsymbol{\mu}}^+_{a+2,y}$ = `TensorDecomposition`$(\hat{\mathbf{T}}^+)$
    $\hat{\boldsymbol{\mu}}^-_{a,y}, \hat{\boldsymbol{\mu}}^-_{a+1,y}, \hat{\boldsymbol{\mu}}^-_{a+2,y}$ = `TensorDecomposition`$(\hat{\mathbf{T}}^-)$
end for

▷ Step 3: Infer Pseudo-Labels
$\tilde{Z}^{(i)} = \tilde{z}_i \sim Y | \lambda_a = \lambda_a^{(i)}, \dots \lambda_m = \lambda_m^{(i)}; \hat{\boldsymbol{\theta}}$

return $\{\tilde{z}_i\}_{i=1}^n$

We cannot directly estimate these parameters from observations of $\boldsymbol{\lambda}_a$, due to the fact that $\mathbf{y}$ is not observed. However, we can observe various moments of the outputs of the LFs. In particular we can observe tensors of outer products of LF triplets:

$$\mathbf{T}^+ := \mathbb{E}[\boldsymbol{\lambda}^+_a \otimes \boldsymbol{\lambda}^+_b \otimes \boldsymbol{\lambda}^+_c] = \sum_{y\in S_Y} w_y \boldsymbol{\mu}^+_{a,y} \otimes \boldsymbol{\mu}^+_{b,y} \otimes \boldsymbol{\mu}^+_{c,y} \quad \text{and} \quad \hat{\mathbf{T}}^+ := \frac{1}{n}\sum_{i=1}^n \boldsymbol{\lambda}^+_{a,i} \otimes \boldsymbol{\lambda}^+_{b,i} \otimes \boldsymbol{\lambda}^+_{c,i}. \tag{5}$$

Here w_y are the mixture probabilities (prior probabilities of Y) and $S_Y = \{y : w_y > 0\}$. We can similarly define $\mathbf{T}^-$ and $\hat{\mathbf{T}}^-$. This allows us to obtain estimates $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ using the tensor decomposition algorithm of [AGH+14] with minor modifications arising from the fact that we work with pseudo-Euclidean rather than Euclidean space. The overall approach is shown in Algorithm 1. We have one key assumption,

Assumption 1. *Assume that the support of P_Y, i.e., $k = |\{y : w_y > 0\}|$ satisfies $k \le d$.*

Our first theoretical result shows that we have near-consistency in estimating the mean parameters in (3). We use standard notation $\tilde{\mathcal{O}}$ that is $\mathcal{O}$ but ignoring the logarithmic factors.

Theorem 1. *Let $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ be the estimates of $\boldsymbol{\mu}^+_{a,y}, \boldsymbol{\mu}^-_{a,y}$ returned by Algorithm 1 with input $\hat{\mathbf{T}}^+, \hat{\mathbf{T}}^-$ constructed using isometric pseudo-Euclidean embeddings (in $\mathbb{R}^{d^+,d^-}$) of n (suff. large) i.i.d observations drawn from the models in 3, $k = |S_Y|$, then $\exists$ constant $C_0 > 0$ such that with high probability $\forall a \in [m]$ and $y \in S_Y$,*

$$|\theta_a - \hat{\theta}_a| \le C_0 \Big| \mathbb{E}_{\lambda_a|y}[d^2_{\mathcal{Y}}(\lambda_a, y)] - \hat{\mathbb{E}}_{\lambda_a|y}[d^2_{\mathcal{Y}}(\lambda_a, y)] \Big| \le \epsilon(d^+) + \epsilon(d^-),$$

where

$$\epsilon(d) := \begin{cases} \tilde{\mathcal{O}}\Big(k\sqrt{\frac{d}{n}}\Big) + \tilde{\mathcal{O}}\Big(\frac{\sqrt{k}}{d}\Big) & \text{if } \sigma^2 = \Theta(1), \\ \tilde{\mathcal{O}}\Big(\sqrt{\frac{k}{n}}\Big) + \tilde{\mathcal{O}}\Big(\frac{\sqrt{k}}{d}\Big) & \text{if } \sigma^2 = \Theta(\frac{1}{d}). \end{cases} \tag{6}$$

Now we interpret Theorem 1. We first note that it is a nearly direct application of [AGJ14]. There are two noise cases for σ. In the high-noise case, σ is independent of dimension d (and thus $|\mathcal{Y}|$). Intuitively, this means the average distance balls around each LF begin to overlap as the number of points grows—explaining the multiplicative k term. If the noise scales down as we add more embedded points, this problem is removed, as in the low-noise case. In both cases, the second error term comes from using the algorithm of [AGH$^+$14] and is independent of the sampling error. Since $k = \Theta(d)$, this term goes down with d. The first error term is due to sampling noise and goes to zero in the number of samples n. Note the tradeoffs of using the embeddings. If we used one-hot encoding, $d = |\mathcal{Y}|$, and in the high-noise case, we would pay a very heavy cost for $\sqrt{d/n}$. However, while sampling error is minimized when using a very small d, we pay a cost in the second error term. This leads to a tradeoff in selecting the appropriate embedding dimension.

We briefly sketch the proof. We show that $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ are accurate estimates of $\boldsymbol{\mu}^+_{a,y}$ and $\boldsymbol{\mu}^+_{a,y}$, leading to accurate estimates of $\mathbb{E}_{\boldsymbol{\lambda}_a|\mathbf{y}}[d^2_\phi(\boldsymbol{\lambda}_a, \mathbf{y})]$. The first of these is done by adapting the requirements from [AGJ14] and running the result twice for the two embedding components.

4 Generalization Error for SP in Finite Metric Spaces

We have access to labeling function outputs $\lambda^{(i)}_a, \ldots, \lambda^{(i)}_m$ and noise rate estimates $\hat{\theta}_a$. How can we use these to replace true (but unobserved) labels y in (2)? Our approach is based on [NDRT13, vRW18]. These works deal with noisy labels by modifying the underlying loss function. Analogously, we show that it is possible to modify (2) in such a way that the generalization guarantee is nearly preserved.

4.1 Prediction with Pseudolabels

First, we construct the posterior distribution $P_{\hat{\boldsymbol{\theta}}}(Y = y|\lambda)$. We use our estimated noise model $P_{\hat{\boldsymbol{\theta}}}(\lambda|Y)$ and the prior $P(Y = y)$, which we assume is known. We create pseudo-labels for each data point by drawing a random sample from the posterior distribution conditioned on the output of labeling functions:

$$\tilde{Z}^{(i)} = \tilde{z}_i \sim Y|\lambda_a = \lambda^{(i)}_a, \ldots \lambda_m = \lambda^{(i)}_m; \hat{\boldsymbol{\theta}}. \tag{7}$$

We thus observe $(x_1, \tilde{z}_1), \ldots (x_n, \tilde{z}_n)$ where $\tilde{z}_i$. To overcome the effect of noise we create a perturbed version of the distance function using the noise rates, generalizing [NDRT13]. Let $\mathcal{Y}^m$ denote the m-fold Cartesian product of $\mathcal{Y}$ and let $\Lambda_u = (\lambda^{(u)}_1, \ldots, \lambda^{(u)}_m)$ denote its u^{th} entry.

$$\mathbf{P}_{ij} = P_{\boldsymbol{\theta}}(\tilde{Z} = y_i|Y = y_j) = \sum_{u=1}^{|\mathcal{Y}^m|} P_{\boldsymbol{\theta}}(Y = y_i|\Lambda = \Lambda^{(u)}) \cdot P_{\boldsymbol{\theta}}(\Lambda = \Lambda^{(u)}|Y = y_j). \tag{8}$$

Similarly define $\mathbf{Q}_{ij} = P_{\hat{\boldsymbol{\theta}}}(\tilde{Z} = y_i|Y = y_j)$ as above but using the estimated parameters $\hat{\boldsymbol{\theta}}$ instead. Note that $\mathbf{P}$ is the true noise distribution introduced by running the inference procedure with the true parameters $\boldsymbol{\theta}$ of the noise model and $\mathbf{Q}$ is an approximation of the noise distribution obtained by performing inference with the *estimated* parameters $\hat{\boldsymbol{\theta}}$.

With this terminology, we can define the perturbed version of the distance function and a corresponding replacement of (2):

$$\tilde{d}_q(T, \tilde{Y} = y_i) := \sum_{j=1}^{k} (\mathbf{Q}^{-1})_{ij} d^2_{\mathcal{Y}}(T, Y = y_j) \quad \forall y_i \in \mathcal{Y} \tag{9}$$

$$\tilde{F}_q(x, y) := \frac{1}{n} \sum_{i=1}^{n} \alpha_i(x) \tilde{d}_q(y, \tilde{z}_i) \qquad \hat{f}_q(x) = \arg\min_{y \in \mathcal{Y}} \tilde{F}_q(x, y) \tag{10}$$

Similarly define, $\tilde{d}_p, \tilde{F}_p, \hat{f}_p$ using the true noise distribution $\mathbf{P}$. It can be easily shown that the perturbed distance function $\tilde{d}_p$ is an unbiased estimator of the true distance function. However we do

not know the true noise distribution $\mathbf{P}$ hence we cannot use it for prediction. Instead we use $\tilde{d}_q$ based on the estimated noise distribution $\mathbf{Q}$. Note that $\tilde{d}_q$ is no longer an unbiased estimator w.r.t to the true noise distribution. However, we can control its bias as a function of the parameter recovery error bound in Theorem 1.

4.2 Bounding the Generalization Error

A natural question to ask is whether the predictor $\hat{f}_q$ will generalize to new data. More concretely, what can we say about the excess risk $R(\hat{f}_q) - R(f^*)$? Note that compared to the prediction based on clean labels, there are two additional sources of error. One is the noise in the labels (i.e., even if we know the true $\mathbf{P}$, the quality of the pseudolabels is imperfect). The other is our estimation procedure for the noise distribution. We must address both sources of error.

We make the following assumptions on the minimum and maximum singular values $\sigma_{\min}(\mathbf{P})$, $\sigma_{\max}(\mathbf{P})$ and the condition number $\kappa(\mathbf{P})$ of true noise matrix $\mathbf{P}$ and the function F. Additional detail is provided in the Appendix.

Assumption 2. *(Noise model is not arbitrary) Assume that the true parameters $\boldsymbol{\theta}$ are such that $\sigma_{\min}(\mathbf{P}) > 0$, and the condition number $\kappa(\mathbf{P})$ is sufficiently small.*

Assumption 3. *(Normalized features) Assume that $|\alpha(x)| \leq 1 \, \forall x \in \mathcal{X}$.*

Assumption 4. *(Proxy strong convexity) Assume that the function F in* (2) *satisfies the following property with some $\beta > 0$, i.e. as we move away from the minimizer of F, the function increases and the rate of increase is proportional to the distance between the points.*

$$F\big(x, f(x)\big) \geq F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}\big(f(x), \hat{f}(x)\big) \qquad \forall x \in \mathcal{X}, \forall f \in \mathcal{F}. \tag{11}$$

With these assumptions, we provide a generalization result for prediction with pseudolabels,

Theorem 2. *(Generalization Error) Let $\hat{f}$ be the minimizer as defined in* (2) *over the clean labels and let $\hat{f}_q$ (defined in* (10)*) be the minimizer over the noisy labels obtained from weak supervision inference in Algorithm 1. Suppose assumptions 2,3,4 hold. Then there exist constants $C_1, C_2 > 0$ dependent on $\sigma_{\max}(\mathbf{P}), \sigma_{\min}(\mathbf{P})$ and k such that w.h.p.,*

$$R(\hat{f}_q) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}) + \tilde{\mathcal{O}}\Big(\frac{C_1}{\beta} n^{-\frac{1}{2}}\Big) + \tilde{\mathcal{O}}\Big(\frac{C_2}{\beta}(\epsilon(d^+) + \epsilon(d^-))\Big). \tag{12}$$

Implications and Tradeoffs: We interpret each term in the bound. The first term is present even with access to the clean labels and hence unavoidable. The second term is the additional error we incur if we learn with the knowledge of the true noise distribution. The third term is due to the use of the estimated noise model. It is dominated by the noise rate recovery result in Theorem 1. If the third term goes to 0, i.e. if we have perfect recovery of the true noise, then we obtain the rate $\mathcal{O}(n^{-1/4})$, the same as in the case of access toclean labels. The third term is introduced by our noise rate recovery algorithm and has two terms: one dominated by $\tilde{\mathcal{O}}(n^{-1/2})$ and the other on $\tilde{\mathcal{O}}(\sqrt{k}/d)$ (see discussion of Theorem 1). Thus we see that we only pay an extra additive factor $\mathcal{O}(\sqrt{k}/d)$ in the excess risk when using pseudolabels. This extra term is negligible for large $d \gg k$.

We briefly sketch the proof. The result follows by first bounding the risk gap between the model learned with the knowledge of noise distribution and the model learned with clean labels i.e. $|R(\hat{f}_p) - R(\hat{f})|$ and then combining it with risk gap between $|R(\hat{f}_q) - R(\hat{f}_p)|$. To obtain the first bound, we use the assumptions (2,3) to argue that F_p is a good approximation of F, i.e. the gap between them is uniformly bounded over all x, y. This fact allows us to show that the minimizers of F_p and F (i.e. $\hat{f}_p$, $\hat{f}$) cannot be far off if the assumption 4 holds. To show the latter, we follow a similar argument to first show a uniform convergence bound for F_p, F_q using the noise rate recovery result and then show a proximity result for their minimizers $\hat{f}_p, \hat{f}_q$ and finally using triangle inequality argue that $\hat{f}_q$ cannot be too far from f^* if, $\hat{f}_p$ and $\hat{f}$ are close to f^*.

5 Manifold-Valued Label Spaces: Noise Recovery and Generalization

We introduce a simple recovery method for weak supervision in constant-curvature Riemannian manifolds. First we briefly introduce some background notation on these spaces, then provide our estimator and consistency result, then the downstream generalization result. Finally, we discuss extensions to symmetric Riemannian manifolds, an even more general class of spaces.

Background on Riemannian manifolds The following is necessarily a very abridged background; more detail can be found in [Lee00, Tu11]. A smooth manifold M is a space where each point is located in a neighborhood diffeomorphic to $\mathbb{R}^d$. Attached to each point $p \in \mathcal{M}$ is a *tangent space* T_pM; each such tangent space is a d-dimensional vector space enabling the use of calculus.

A Riemannian manifold equips a smooth manifold with a Riemannian metric: a smoothly-varying inner product $\langle \cdot, \cdot \rangle_p$ at each point p. This tool allows us to compute angles, lengths, and ultimately, distances $d_{\mathcal{M}}(p, q)$ between points on the manifold as shortest-path distances. These shortest paths are called geodesics and can be parametrized as curves $\gamma(t)$, where $\gamma(0) = p$, or by tangent vectors $v \in T_pM$. The exponential map operation $\exp : T_p\mathcal{M} \to \mathcal{M}$ takes tangent vectors to manifold points. It enables switching between these tangent vectors: $\exp_p(v) = q$ implies that $d_{\mathcal{M}}(p, q) = \|v\|$.

Invariant Our first contribution is a simple invariant that enables us to recover the error parameters. Note that the finite metric-space case is insufficient: the support is infinite. Nor do we need an embedding—we have a continuous representation as-is. Instead, we propose a simple idea based on the law of cosines. Essentially, on average, the geodesic triangle formed by the latent variable $y \in \mathcal{M}$ and two observed LFs λ^a, λ^b, is a right triangle. This means it can be characterized by the (Riemannian) version of the Pythagorean theorem:

Lemma 1. *For $\mathcal{Y} = \mathcal{M}$, a hyperbolic manifold, $y \sim P$ for some distribution P on $\mathcal{M}$ and labeling functions λ^a, λ^b drawn from* (3)*,*

$$\mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^a, \lambda^b) = \mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^b, y) \mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^b, y),$$

while for $\mathcal{Y} = \mathcal{M}$ a spherical manifold,

$$\mathbb{E} \cos d_{\mathcal{Y}}(\lambda^a, \lambda^b) = \mathbb{E} \cos d_{\mathcal{Y}}(\lambda^b, y) \mathbb{E} \cos d_{\mathcal{Y}}(\lambda^b, y).$$

These invariants enable us to easily learn by forming a triplet system. Suppose we construct the equation in Lemma 1 for three pairs of labeling functions. The resulting system can be solved to express $\mathbb{E}[\cosh(d_{\mathcal{Y}}(\lambda^a, y))]$ in terms of $\mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^a, \lambda^b)), \mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^a, \lambda^c)), \mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^b, \lambda^c))$. Specifically,

$$\mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^a, y)) = \sqrt{\frac{\mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^a, \lambda^b) \mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^a, \lambda^c)}{(\mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^b, \lambda^c))^2}}.$$

Note that we can estimate $\hat{\mathbb{E}}$ via the empirical versions of terms on the right , as these are based on observable quantities. This is a generalization of the binary case in [FCS+20] and the Gaussian (Euclidean) case in [SLV+22] to hyperbolic manifolds. A similar estimator can be obtained for spherical manifolds by replacing $\cosh$ with $\cos$.

Using this tool, we can obtain a consistent estimator for θ^a for each of $a = 1, \ldots, m$. Let C_0 satisfy $\mathbb{E}|\hat{\mathbb{E}} \cosh(d_{\mathcal{Y}}(\lambda^a, \lambda^b)) - \mathbb{E} \cosh(d_{\mathcal{Y}}(\lambda^a, \lambda^b))| \geq C_0 \mathbb{E}|\hat{\mathbb{E}} d^2_{\mathcal{Y}}(\lambda^a, \lambda^b)) - \mathbb{E} d^2_{\mathcal{Y}}(\lambda^a, \lambda^b)|$; that is, C_0 reflects the pushforward of concentration between the distributions $\cosh(d)$ and d^2. Then,

Theorem 3. *Let $\mathcal{M}$ be a hyperbolic manifold. Fix $0 < \delta < 1$ and let $\Delta(\delta) = \min_\rho Pr(\forall i, d_{\mathcal{Y}}(\lambda^a(i), \lambda^b(i) \leq \rho)) \geq 1 - \delta$. Then, there exists a constant C_1 so that with probability at least $1 - \delta$,*

$$\mathbb{E}|\hat{\mathbb{E}} d^2_{\mathcal{Y}}(\lambda^a, y)) - \mathbb{E} d^2_{\mathcal{Y}}(\lambda^a, y)| \leq \frac{C_1 \cosh(\Delta(\delta))^{3/2}}{C_0 \sqrt{2n}}.$$

As we hoped, our estimator is consistent. Note that we pay a price for a tighter bound: $\Delta(\delta)$ is large for smaller probability δ. It is possible to estimate the size of $\Delta(\delta)$ (more generally, it is a function of the curvature). We provide more details in the Appendix.

Next, we adapt the downstream model predictor (2) in the following way. Let $\hat{\mu}_a^2 = \hat{\mathbb{E}}[d_{\mathcal{Y}}^2(\lambda^a, y)]$. Let $\beta = [\beta_1, \ldots, \beta_m]^T$ be such that $\sum_a \beta_a = 1$ and β minimizes $\sum_a \beta_a^2 \hat{\mu}_a^2$. Then, we set

$$\tilde{f}(x) = \arg\min_{y \in \mathcal{Y}} \frac{1}{n} \sum_{i=1}^{n} \alpha_i(x) \sum_{a=1}^{m} \beta_a^2 d_{\mathcal{Y}}^2(y, \lambda_{a,i}).$$

We simply replace each of the true labels with a combination of the labeling functions. With this, we can state our final result. First, we introduce our assumptions.

Let $q = \arg\min_{z \in \mathcal{Y}} \mathbb{E}[\alpha(x)(y) d_{\mathcal{Y}}^2(z, y)]$, where the expectation is taken over the population level distribution and $\alpha(x)(y)$ denotes the kernel at y.

Assumption 5. *(Bounded Hugging Function c.f. [Str20]) Let q be defined as above. For all $a, b \in \mathcal{M}$, the hugging function at q is given by $k_q^b(a) = 1 - (\|\log_q(a) - \log_q(b)\|^2 - d_{\mathcal{Y}}^2(a,b))/d_{\mathcal{Y}}^2(q,b)$. We assume that $k_q^b(a)$ is lower bounded by $k_{\min}$.*

Assumption 6. *(Kernel Symmetry) We assume that for all x and all $v \in T_q\mathcal{M}$, $\alpha(x)(\exp_q(v)) = \alpha(x)(\exp_q(-v))$.*

The first condition provides control on how geodesic triangles behave; it relates to the curvature. We provide more details on this in the Appendix. The second assumption restricts us to kernels symmetric about the minimizers of the objective F. Finally, suppose we draw (x, y) and (x', y') independently from P_{XY}. Set $\sigma_o^2 = \alpha(x)(y)\mathbb{E}d_{\mathcal{Y}}^2(y, y')$.

Theorem 4. *Let $\mathcal{M}$ be a complete manifold and suppose the assumptions above hold. Then, there exist constants C_3, C_4*

$$\mathbb{E}[d_{\mathcal{Y}}^2(\hat{f}(x), \tilde{f}(x))] \leq \frac{C_3 \sigma_o^2}{n k_{\min}} + \frac{C_4 \sum_{a=1}^{m} \beta_a^2 \hat{\mu}_a^2}{m n k_{\min}}.$$

Note that as both m and n grow, as long as our worst-quality LF has bounded variance, our estimator of the true predictor is consistent. Moreover, we also have favorable dependence on the noise rate. This is because the only error we incur is in computing suboptimal β coefficients. We comment on this suboptimality in the Appendix.

A simple corollary of Theorem 5 provides the generalization guarantees we sought,

Corollary 1. *Let $\mathcal{M}$ be a complete manifold and suppose the assumptions above hold. Then, with high probability,*

$$R(\tilde{f}) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}).$$

Extensions to Other Manifolds First, we note that all of our approaches almost immediately lift to products of constant-curvature spaces. For example, we have that $\mathcal{M}_1 \times \mathcal{M}_2$ has metric $d_{\mathcal{Y}}^2(p, q) = d_{\mathcal{M}_1}^2(p_1, q_1) + d_{\mathcal{M}_2}^2(p_2, q_2)$, where p_i, q_i are the projections of p, q onto the ith component.

We can go beyond products of constant-curvature spaces as well. To do so, we can build generalizations of the law of cosines (as needed for the invariance in Lemma 1). For example, it is possible to do for symmetric Riemannian manifolds using the tools in [AH91].

6 Conclusion

We studied the theoretical properties of weak supervision applied to structured prediction. Our focus was on two general scenarios: label spaces that are finite metric spaces or continuous spaces given by constant-curvature manifolds. In both scenarios, we introduced ways to estimate the noise rates of labeling functions, achieving consistency or near-consistency. Using these estimators, we established that, after suitable modifications, downstream structured prediction models maintain their generalization guarantees. Future directions include extending these results to even more general manifolds and removing some of the assumptions that limit our results to particular models.

References

[AGH+14] Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *Journal of Machine Learning Research*, 15:2773–2832, 2014.

[AGJ14] Animashree Anandkumar, Rong Ge, and Majid Janzamin. Sample complexity analysis for learning overcomplete latent variable models through tensor methods. *arXiv preprint arXiv:1408.0553*, 2014.

[AH91] Helmer Aslaksen and Hsueh-Ling Huynh. Laws of trigonometry in symmetric spaces. *Geometry from the Pacific Rim*, 1991.

[BHS+07] Gükhan H. Bakir, Thomas Hofmann, Bernhard Schölkopf, Alexander J. Smola, Ben Taskar, and S. V. N. Vishwanathan. *Predicting Structured Data (Neural Information Processing)*. The MIT Press, 2007.

[BS16] R.B. Bapat and Sivaramakrishnan Sivasubramanian. Squared distance matrix of a tree: Inverse and inertia. *Linear Algebra and its Applications*, 491:328–342, 2016.

[CRR16] Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi. A consistent regularization approach for structured prediction. In *Advances in Neural Information Processing Systems 30 (NIPS 2016)*, volume 30, 2016.

[Dem92] James Demmel. The component-wise distance to the nearest singular matrix. *SIAM Journal on Matrix Analysis and Applications*, 13(1):10–19, 1992.

[DRS+20] Jared A. Dunnmon, Alexander J. Ratner, Khaled Saab, Nishith Khandwala, Matthew Markert, Hersh Sagreiya, Roger Goldman, Christopher Lee-Messer, Matthew P. Lungren, Daniel L. Rubin, and Christopher Ré. Cross-modal data programming enables rapid medical machine learning. *Patterns*, 1(2), 2020.

[FCS+20] Daniel Y. Fu, Mayee F. Chen, Frederic Sala, Sarah M. Hooper, Kayvon Fatahalian, and Christopher Ré. Fast and three-rious: Speeding up weak supervision with triplet methods. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, 2020.

[Gol85] Lev Goldfarb. A new approach to pattern recognition. pages 241–402, 1985.

[GS19] Colin Graber and Alexander Schwing. Graph structured prediction energy networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2019)*, volume 33, 2019.

[KAG18] Anna Korba and Florence d'Alché-Buc Alexandre Garcia. A structured prediction approach for label ranking. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2018)*, volume 32, 2018.

[KL15] Volodymyr Kuleshov and Percy S Liang. Calibrated structured prediction. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, 2015.

[KW78] J.B. Kruskal and M. Wish. *Multidimensional Scaling*. Sage Publications, 1978.

[Lee00] John M. Lee. *Introduction to Smooth Manifolds*. Springer, 2000.

[NDRT13] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'13, page 1196–1204, 2013.

[PM19] Alexander Petersen and Hans-Georg Müller. Fréchet regression for random objects with euclidean predictors. *Annals of Statistics*, 47(2):691–719, 2019.

[RBE+18] Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the 44th International Conference on Very Large Data Bases (VLDB)*, Rio de Janeiro, Brazil, 2018.

[RCMR18] Alessandro Rudi, Carlo Ciliberto, GianMaria Marconi, and Lorenzo Rosasco. Manifold structured prediction. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2018)*, volume 32, 2018.

[RHD+19] A. J. Ratner, B. Hancock, J. Dunnmon, F. Sala, S. Pandey, and C. Ré. Training complex models with multi-task weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Honolulu, Hawaii, 2019.

[RNGS20] Christopher Ré, Feng Niu, Pallavi Gudipati, and Charles Srisuwananukorn. Overton: A data system for monitoring and improving machine-learned products. In *Proceedings of the 10th Annual Conference on Innovative Data Systems Research*, 2020.

[RSW+16] A. J. Ratner, Christopher M. De Sa, Sen Wu, Daniel Selsam, and C. Ré. Data programming: Creating large training sets, quickly. In *Proceedings of the 29th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain, 2016.

[SLB20] Esteban Safranchik, Shiying Luo, and Stephen Bach. Weakly supervised sequence tagging from noisy rules. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, pages 5570–5578, Apr. 2020.

[SLV+22] Changho Shin, Winfred Li, Harit Vishwakarma, Nicholas Carl Roberts, and Frederic Sala. Universalizing weak supervision. In *International Conference on Learning Representations*, 2022.

[Str20] Austin J. Stromme. *Wasserstein Barycenters: Statistics and Optimization*. MIT, 2020.

[Tu11] Loring W. Tu. *An Introduction to Manifolds*. Springer, 2011.

[vRW18] Brendan van Rooyen and Robert C. Williamson. A theory of learning with corrupted labels. *Journal of Machine Learning Research*, 18(228):1–50, 2018.

[ZS16] Hongyi Zhang and Suvrit Sra. First-order methods for geodesically convex optimization. In *Conference on Learning Theory, COLT 2016*, 2016.

Checklist

1. Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
2. Did you describe the limitations of your work? [Yes]
3. Did you discuss any potential negative societal impacts of your work? [N/A]
4. Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
5. Did you state the full set of assumptions of all theoretical results? [Yes]
6. Did you include complete proofs of all theoretical results? [Yes] See appendix

Appendix

The Appendix is organized as follows. First, we provide a glossary that summarizes the notation we use throughout the paper. Afterwards, we provide the proofs for the finite-valued metric space cases. We continue with the proofs and additional discussion for the manifold-valued label spaces. Finally, we give some additional explanations for pseudo-Euclidean spaces.

A Glossary

The glossary is given in Table 1 below.

Symbol	Definition
$\mathcal{X}$	feature space
$\mathcal{Y}$	label metric space
S_Y	support of prior distribution on true labels
$d_{\mathcal{Y}}$	label metric (distance) function
$x_1, x_2, \ldots, x_n$	unlabeled datapoints from $\mathcal{X}$
$y_1, y_2, \ldots, y_n$	latent (unobserved) labels from $\mathcal{Y}$
$s_1, s_2, \ldots, s_m$	labeling functions / sources
$\lambda_1, \lambda_2, \ldots, \lambda_m$	output of labeling functions (LFs)
$\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2, \ldots, \boldsymbol{\lambda}_m$	pseudo-Euclidean embeddings of LFs outputs
$\lambda_a^{(i)}$	output of ath LF on ith data point x_i
$\boldsymbol{\lambda}_a^{(i)}$	pseudo-Euclidean Embedding of output of ath LF on ith data point x_i
n	number of data points
m	number of LFs
k	size of the support of prior on $\mathcal{Y}$ i.e. $k = \lvert S_Y \rvert$
r	size of $\mathcal{Y}$ for the finite case
$\lambda_a^{(i)}$	output of ath labeling function applied to ith sample x_i
$\theta_a, \hat{\theta}_a$	true and estimated canonical parameters of model in (3)
$\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}$	true and estimated canonical parameters arranged as vectors.
$\mathbb{E}[d^2_{\mathcal{Y}}(\lambda_a, y)]$	mean parameters in (3)
g	pseudo-Euclidean embedding mapping
$\mathbf{P}$	true noise model $P_{ij} = P_{\boldsymbol{\theta}}(\tilde{Y} = y_i \lvert Y = y_j)$ with true parameters $\boldsymbol{\theta}$
$\mathbf{Q}$	estimated noise model with parameters $\hat{\boldsymbol{\theta}}$, $Q_{ij} = P_{\hat{\boldsymbol{\theta}}}(\tilde{Y} = y_i \lvert Y = y_j)$
Λ	a random element in $\mathcal{Y}^m$ the m-fold Cartesian product of $\mathcal{Y}$.
$\Lambda^{(u)}$	uth element in $\mathcal{Y}^m$
$\boldsymbol{\mu}^+_{a,y}, \boldsymbol{\mu}^-_{a,y}$	means of distributions in (4) corresponding to $\mathbb{R}^{d^+}, \mathbb{R}^{d^-}$
$\epsilon(d^+), \epsilon(d^-)$	error in recovering the mean parameters, (6)
σ	noise variance in (4)
$F(x, y)$	the score function in (2) with true labels
$\tilde{F}_p(x, y), \tilde{F}_q(x, y)$	the score function in (10) with noisy labels from distributions $\mathbf{P}$ and $\mathbf{Q}$
$\hat{f}$	minimizer of F defined in (2)
$\hat{f}_p, \hat{f}_q$	minimizers of $\tilde{F}_p, \tilde{F}_q$ as defined in (2)
$\sigma_{\max}(\mathbf{P})$	maximum singular value of $\mathbf{P}$
$\sigma_{\min}(\mathbf{P})$	minimum singular value of $\mathbf{P}$
$\kappa(\mathbf{P})$	the condition number of matrix $\mathbf{P}$

Table 1: Glossary of variables and symbols used in this paper.

We introduce results leading to the proofs of the theorems for the finite-valued metric space case.

Lemma 2. *([AGJ14]) Let $\hat{\mathbf{T}}^+, \hat{\mathbf{T}}^-$ be the third order observed moments for labeling functions triplet (a, b, c), as defined in 5 over n (suff. large) i.i.d observations drawn from models in equation 4, and $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^+_{b,y}, \hat{\boldsymbol{\mu}}^+_{c,y}$ and $\hat{\boldsymbol{\mu}}^-_{a,y}, \hat{\boldsymbol{\mu}}^-_{b,y}, \hat{\boldsymbol{\mu}}^-_{c,y}$ be the estimated parameters returned by the algorithm 1. Let*

$\epsilon(d)$ *be defined as above in equation 6, then the following holds with high probability for all triplets* (a, b, c) *of labeling functions,*

$$||\boldsymbol{\mu}^+_{s,y} - \hat{\boldsymbol{\mu}}^+_{s,y}||_2 \leq \mathcal{O}(\epsilon(d^+)) \quad \textit{and} \quad ||\boldsymbol{\mu}^-_{s,y} - \hat{\boldsymbol{\mu}}^-_{s,y}||_2 \leq \mathcal{O}(\epsilon(d^-)) \quad \forall s \in (a, b, c) \, \forall y \in \mathcal{Y} \tag{13}$$

Proof. The result in [AGJ14] is in terms of the following distance function,

$$\text{dist}(\mathbf{u}, \mathbf{v}) = \sup_{\mathbf{z} \perp \mathbf{u}} \frac{\langle \mathbf{z}, \mathbf{v} \rangle}{||\mathbf{z}||_2 ||\mathbf{v}||_2} = \sup_{\mathbf{z} \perp \mathbf{v}} \frac{\langle \mathbf{z}, \mathbf{u} \rangle}{||\mathbf{z}||_2 ||\mathbf{u}||_2}.$$

The proof follows by translating the result to the euclidean distance. for $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ with $||\mathbf{u}||, ||\mathbf{v}|| = 1$,

$$\min_{z \in \{-1, +1\}} ||z\mathbf{u} - \mathbf{v}||_2 \leq \sqrt{2}\, \text{dist}(\mathbf{u}, \mathbf{v}).$$

This notion of distance is oblivious to sign recovery. However if the sign recovery is possible then we have,

$$||\mathbf{u} - \mathbf{v}||_2 \leq \sqrt{2}\, \text{dist}(\mathbf{u}, \mathbf{v}).$$

We are assuming that sign recovery is possible, (in the worst case by doing brute-force over the the signs for each LF.) And further with appropriate normalization we have $||\boldsymbol{\mu}^+_a|| = 1, ||\hat{\boldsymbol{\mu}}^+_a|| = 1$ and $||\boldsymbol{\mu}^-_a|| = 1, ||\hat{\boldsymbol{\mu}}^-_a|| = 1$,

This gives us

$$||\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}||_2 \leq \mathcal{O}(\text{dist}(\boldsymbol{\mu}^+_{a,y}, \boldsymbol{\mu}^-_{a,y})) \leq \mathcal{O}(\epsilon(d^+)).$$

and similarly for $\boldsymbol{\mu}^-_{a,y}$. Further with n, d be suff. large such that $\epsilon(d^+), \epsilon(d^-) \leq 1$, then the result holds for squared distances. □

Theorem 1. *Let* $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ *be the estimates of* $\boldsymbol{\mu}^+_{a,y}, \boldsymbol{\mu}^-_{a,y}$ *returned by Algorithm 1 with input* $\hat{\mathbf{T}}^+, \hat{\mathbf{T}}^-$ *constructed using isometric pseudo-Euclidean embeddings (in* $\mathbb{R}^{d^+, d^-}$*) of* n *(suff. large) i.i.d observations drawn from the models in 3,* $k = |S_Y|$*, then* $\exists$ *constant* $C_0 > 0$ *such that with high probability* $\forall a \in [m]$ *and* $y \in S_Y$*,*

$$|\theta_a - \hat{\theta}_a| \leq C_0 \Big| \mathbb{E}_{\lambda_a | y}[d^2_{\mathcal{Y}}(\lambda_a, y)] - \hat{\mathbb{E}}_{\lambda_a | y}[d^2_{\mathcal{Y}}(\lambda_a, y)] \Big| \leq \epsilon(d^+) + \epsilon(d^-),$$

where

$$\epsilon(d) := \begin{cases} \tilde{\mathcal{O}}\Big(k\sqrt{\frac{d}{n}}\Big) + \tilde{\mathcal{O}}\Big(\frac{\sqrt{k}}{d}\Big) & \textit{if } \sigma^2 = \Theta(1), \\ \tilde{\mathcal{O}}\Big(\sqrt{\frac{k}{n}}\Big) + \tilde{\mathcal{O}}\Big(\frac{\sqrt{k}}{d}\Big) & \textit{if } \sigma^2 = \Theta(\frac{1}{d}). \end{cases} \tag{6}$$

Proof. Using the tensor decomposition result from lemma 2 we get, estimate $\hat{\boldsymbol{\mu}}_{a,y}$ such that

$$||\hat{\boldsymbol{\mu}}^+_{a,y} - \boldsymbol{\mu}^+_{a,y}||^2_2 \leq \mathcal{O}(\epsilon(d^+)) \quad \text{and} \quad ||\hat{\boldsymbol{\mu}}^-_{a,y} - \boldsymbol{\mu}^-_{a,y}||^2_2 \leq \mathcal{O}(\epsilon(d^-)).$$

Using the definition of euclidean distance and the fact that $\mathbb{E}[\boldsymbol{\lambda}_a[i]^2] - \boldsymbol{\mu}_{a,y}[i]^2 = \sigma_i^2$, we get

$$\begin{aligned} \mathbb{E}_{\boldsymbol{\lambda}^+_a | \mathbf{y}}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y})] &= \mathbb{E}_{\boldsymbol{\lambda}^+_a | \mathbf{y}}\Big[||\boldsymbol{\lambda}^+_a||^2_2 + ||\mathbf{y}^+||^2_2 - 2\langle \boldsymbol{\lambda}^+_a, \mathbf{y}^+ \rangle\Big], \\ &= ||\boldsymbol{\mu}^+_{a,y}||^2_2 + \sum_{i=1}^{d^+} \sigma_i^2 + ||\mathbf{y}^+||^2_2 - 2\langle \boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+ \rangle. \end{aligned}$$

Plugging in the estimate of $\boldsymbol{\mu}^+_{a,y}$ we get,

$$\hat{\mathbb{E}}_{\boldsymbol{\lambda}^+_a | \mathbf{y}}[d^2_\phi(\boldsymbol{\lambda}^+_a, y)] = ||\hat{\boldsymbol{\mu}}^+_{a,y}||^2_2 + \sum_{i=1}^{d^+} \sigma_i^2 + ||\mathbf{y}^+||^2_2 - 2\langle \hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+ \rangle.$$

Thus,

$$\begin{aligned}\Big|\mathbb{E}_{\boldsymbol{\lambda}_a^+|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^+,\mathbf{y})] - \hat{\mathbb{E}}_{\boldsymbol{\lambda}_a^+|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^+,\mathbf{y})]\Big| &\leq \Big(||\boldsymbol{\mu}_{a,y}^+||_2^2 - ||\hat{\boldsymbol{\mu}}_{a,y}^+||_2^2\Big) + 2||\mathbf{y}^+||_2\Big(||\boldsymbol{\mu}_{a,y}^+ - \hat{\boldsymbol{\mu}}_{a,y}^+||_2\Big),\\ &\leq \mathcal{O}(\epsilon(d^+)) + \mathcal{O}(\epsilon(d^+)),\\ &= \mathcal{O}(\epsilon(d^+)).\end{aligned}$$

Here we used $||\boldsymbol{\mu}_{a,y}^+ - \hat{\boldsymbol{\mu}}_{a,y}^+||_2 \leq \mathcal{O}(\epsilon(d^+))$ and $||\boldsymbol{\mu}_{a,y}^+||_2, ||\hat{\boldsymbol{\mu}}_{a,y}^+||_2 = 1, ||\mathbf{y}^+||_2 \leq 1$, which allows us to bound,

$$\begin{aligned}||\boldsymbol{\mu}_{a,y}^+||_2^2 - ||\hat{\boldsymbol{\mu}}_{a,y}^+||_2^2 &= \Big\langle\Big(\hat{\boldsymbol{\mu}}_{a,y}^+ - \boldsymbol{\mu}_{a,y}^+\Big), \Big(\hat{\boldsymbol{\mu}}_{a,y}^+ + \boldsymbol{\mu}_{a,y}^+\Big)\Big\rangle,\\ &\leq ||\hat{\boldsymbol{\mu}}_{a,y}^+ - \boldsymbol{\mu}_{a,y}^+||_2 \cdot ||\hat{\boldsymbol{\mu}}_{a,y}^+ + \boldsymbol{\mu}_{a,y}^+||_2,\\ &\leq \mathcal{O}(\epsilon(d^+)).\end{aligned}$$

Doing the same calculations for $\boldsymbol{\lambda}_a^-$, we get

$$\Big|\mathbb{E}_{\boldsymbol{\lambda}_a^-|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^-,\mathbf{y})] - \hat{\mathbb{E}}_{\boldsymbol{\lambda}_a^-|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^-,\mathbf{y})]\Big| \leq \mathcal{O}(\epsilon(d^-)).$$

Thus overall error in mean parameters is

$$\begin{aligned}\Big|\mathbb{E}_{\boldsymbol{\lambda}_a|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a,\mathbf{y})] - \hat{\mathbb{E}}_{\boldsymbol{\lambda}_a|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a,\mathbf{y})]\Big| &\leq \Big|\mathbb{E}_{\boldsymbol{\lambda}_a^+|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^+,\mathbf{y})] - \hat{\mathbb{E}}_{\boldsymbol{\lambda}_a^+|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^+,\mathbf{y})]\Big| +\\ &\qquad \Big|\mathbb{E}_{\boldsymbol{\lambda}_a^-|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^-,\mathbf{y})] - \hat{\mathbb{E}}_{\boldsymbol{\lambda}_a^-|\mathbf{y}}[d_\phi^2(\boldsymbol{\lambda}_a^-,\mathbf{y})]\Big|,\\ &\leq \mathcal{O}(\epsilon(d^+)) + \mathcal{O}(\epsilon(d^-)).\end{aligned}$$

Next, we use a known relation between the mean and the canonical parameters of the exponential model to get the result in terms of the canonical parameters. In particular the result says the following,

$$|\theta_a - \hat{\theta}_a| \leq \frac{1}{e_{\min}(A_a(\theta))}\big|\mathbb{E}[d_{\mathcal{Y}}^2(\lambda_a, y) - \hat{\mathbb{E}}[d_{\mathcal{Y}}^2(\lambda_a, y)]\big|.$$

where $A_a(\theta)$ is the log partition function of the label model in (3) and $e_{\min}(A_a) = \inf_{\theta\in\Theta} \frac{d^2}{d\theta^2} A_a(\theta)$ over the parameter space Θ. For more details see Lemma 8 from [FCS+20] and Theorem 4.3 in [SLV+22]. Letting $C_0 = \max_{a\in[m]} e_{\min}(A_a)$ gives us the result.

□

Finding σ for distribution in (4)

$$\begin{aligned}u(\theta) = \mathbb{E}[d_{\mathcal{Y}}^2(\lambda, y)] = \mathbb{E}[d_\phi^2(\boldsymbol{\lambda},\mathbf{y})] &= \sum_{i=1}^{d^+}\mathbb{E}[(\boldsymbol{\lambda}[i] - \mathbf{y}[i])^2] - \sum_{i=d^++1}^{d}\mathbb{E}[(\boldsymbol{\lambda}[i] - \mathbf{y}[i])^2],\\ &= \Big(\sum_{i=1}^{d^+}\mathbb{E}[(\boldsymbol{\lambda}[i] - \boldsymbol{\mu}_y[i])^2] - (\boldsymbol{\mu}_y[i] - \mathbf{y}[i])^2\Big) -\\ &\qquad \Big(\sum_{i=d^++1}^{d}\mathbb{E}[(\boldsymbol{\lambda}[i] - \boldsymbol{\mu}_y[i])^2] - (\boldsymbol{\mu}_y[i] - \mathbf{y}[i])^2\Big),\\ &= \sum_{i=1}^{d^+}\sigma_i^2 - \sum_{i=d^++1}^{d}\sigma_i^2 + d_\phi^2(\boldsymbol{\mu}_y,\mathbf{y}),\\ &\leq d\sigma_{\max}^2 + d_\phi^2(\boldsymbol{\mu}_y,\mathbf{y}).\end{aligned}$$

$\sigma_{\max}^2 \geq \frac{1}{d}\Big(\mathbb{E}[d_\phi^2(\boldsymbol{\lambda},\mathbf{y})] - d_\phi^2(\boldsymbol{\mu}_y,\mathbf{y})\Big) = \frac{1}{d}\Big(u(\theta) - d_\phi^2(\boldsymbol{\mu}_y,\mathbf{y})\Big)$, $u(\theta)$ is inversely proportional to θ. High θ implies that there is low variance in (3), thus it implies for low variance in (3) we have low $\sigma_{\max}$.

B Proofs for Generalization Error

B.1 When True Noise Distribution is Available

Lemma 3. *Let the distribution $\tilde{Y}|Y$ be given by $\mathbf{P}$ a $k \times k$ transition probability matrix with $\mathbf{P}_{ij} = \mathbb{P}(\tilde{Y} = y_i|Y = y_j)$ and let $\mathbf{P}$ be invertible matrix. Let the pseudo-distance $\tilde{d}_p$ be defined as in equation 9 then,*

$$\mathbb{E}_{\tilde{Y}|Y=y_j}\big[\tilde{d}_p(T, \tilde{Y})\big] = d^2_{\mathcal{Y}}(T, y_j). \tag{14}$$

Proof. It is easy to see it in terms of vectors, denote $\tilde{\mathbf{d}}_p \in \mathbb{R}^k$ with ith entry given by $\tilde{d}_p(T, \tilde{Y} = y_i)$ and similarly define $\tilde{\mathbf{d}}$. Then we can see that $\tilde{\mathbf{d}}$ satisfies the following with $\mathbf{P}$ being a symmetric matrix.

$$\tilde{\mathbf{d}}_p = (\mathbf{P})^{-1}\mathbf{d} \implies \mathbb{E}_{\tilde{Y}|Y}[\tilde{\mathbf{d}}_p] = \mathbf{P}(\mathbf{P})^{-1}\mathbf{d} = \mathbf{d}.$$

□

Lemma 4. *Let F and $\tilde{F}_p$ be defined as in equations* (10) *and 2 over n i.i.d. samples, then the following holds for any $x \in \mathcal{X}, y \in \mathcal{Y}$ w.h.p.*

$$|F(x, y) - \tilde{F}_p(x, y)| \leq \tilde{\mathcal{O}}\Big(\big(\frac{1 + \sigma_{\max}(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\big)\sqrt{\frac{1}{n}}\Big). \tag{15}$$

where $\sigma_{\max}(\mathbf{P}), \sigma_{\min}(\mathbf{P})$ are the maximum and minimum singular values of $\mathbf{P}$.

Proof. Recall the definitions, Let $y_i{}_{i=1}^n$ be the true labels of points $x_i{}_{i=1}^n$ and let the pseudo-label for ith point drawn from noise model $\mathbf{P}$ be $\tilde{y}_i$.

$$F(x, y) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x) d^2_{\mathcal{Y}}(y, y_i), \qquad \tilde{F}_p(x, y) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\tilde{d}_p(y, \tilde{y}_i).$$

$$\begin{aligned}\tilde{F}_p(x, y) - F(x, y) &= \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\Big(\tilde{d}_p(y, \tilde{y}_i) - d^2_{\mathcal{Y}}(y, y_i)\Big), \\ &= \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\xi(y, y_i, \tilde{y}_i).\end{aligned}$$

Here y, y_i are fixed and the randomness is over $\tilde{y}_i$, thus we can think of $\tilde{y}_i$ as random variable $\tilde{Y}_i$ and take the expectation of ξ over the distribution $\mathbf{P}$. We can see that from Lemma 3 we have $\mathbb{E}_{\tilde{Y}\sim\mathbf{P}[:,y_i]}[\xi(y, y_i, \tilde{Y})] = 0$ this implies $\mathbb{E}[\tilde{F}_p(x, y) - F(x, y)] = 0$.

Moreover, $\alpha_i(x) \cdot \xi(y, y_i, \tilde{Y}_i)$ are independent r.v. and $\alpha_i(x) \leq 1$, but we don't know if $\xi(y, y_i, \tilde{Y}_i)$ are bounded. It would be misleading to think of $\tilde{d}_p$ as distance and use the same upper bound as of $d^2_{\mathcal{Y}}$ on it – due to the fact that $\tilde{d}_p$ is obtained by multiplying by inverse of $\mathbf{P}$ and the true distances and the entries of the inverse can have magnitude large than 1. However we can see that ξ are bounded as following as long as the spectral decomposition of $\mathbf{P}$ is not arbitrary,

$$||\tilde{\mathbf{d}}_p - \mathbf{d}||_\infty = ||\mathbf{P}^{-1}\mathbf{d}_p - \mathbf{d}||_\infty \leq ||\mathbf{P}^{-1}||_2||\mathbf{I} - \mathbf{P}||_2||\mathbf{d}||_\infty \leq \frac{1 + \sigma_{max}(\mathbf{P})}{\sigma_{min}(\mathbf{P})} =: c_1.$$

Thus using Hoeffding's inequality,

$$|\tilde{F}_p(x, y) - F(x, y)| \leq \tilde{\mathcal{O}}\Big(c_1\sqrt{\frac{1}{n}}\Big).$$

□

Lemma 5. *Let $\hat{f}$ be the minimizer as defined in equation 2 over the clean labels and let $\hat{f}_p$ (defined in eq. 10) be the minimizer over the noisy labels obtained from conditional distribution $\tilde{Y}|Y$ i.e.* $\mathbf{P}$ *such that lemma 3, 4 hold, and let the risk function be defined as in equation 1, then w.h.p.*

$$d^2_{\mathcal{Y}}\big(\hat{f}_p(x), \hat{f}(x)\big) \leq \tilde{\mathcal{O}}\Big(\frac{c_1}{\beta}\sqrt{\frac{1}{n}}\Big). \tag{16}$$

Proof. Recall the definitions,

$$\hat{f}(x) = \arg\min_{y\in\mathcal{Y}} F(x,y) \qquad \hat{f}_p(x) = \arg\min_{y\in\mathcal{Y}} \tilde{F}_p(x,y)$$

Let $d^2_{\mathcal{Y}}(f_1, f_2) = \sup_{x\in\mathcal{X}} d^2_{\mathcal{Y}}\big(f_1(x), f_2(x)\big)$ and let $\mathcal{B}(\hat{f}, r) = \{f : d^2_{\mathcal{Y}}(\hat{f}, f) \leq r\}$ denote the ball of radius r around $\hat{f}$.

From lemma 4 we know for $t = \tilde{\mathcal{O}}\Big(\big(\frac{1+\sigma_{max}(\mathbf{P})}{\sigma_{min}(\mathbf{P})}\big)\sqrt{\frac{1}{n}}\Big)$,

$$F\big(x, f(x)\big) - t \leq \tilde{F}_p\big(x, f(x)\big) \leq F\big(x, f(x)\big) + t.$$

From assumption 4 we have,

$$F\big(x, f(x)\big) \geq F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}(f(x), \hat{f}(x)).$$

Combining the two we get a lower bound on $\tilde{F}_p$,

$$\tilde{F}_p(x, f(x)) \geq F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}(f(x), \hat{f}(x)) - t.$$

We want to find a suff. large ball around $\hat{f}$ such that the minimizer of $\tilde{F}_p$ does not lie outside this ball. To see this let LB and UB denote the above mentioned lower and upper bounds on $\tilde{F}_p$,

$$\begin{aligned} LB(\tilde{F}_p, f, x) &:= F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}(f(x), \hat{f}(x)) - t. \\ UB(\tilde{F}_p, f, x) &:= F\big(x, f(x)\big) + t. \end{aligned}$$

For $f \in \mathcal{B}(\hat{f}, \frac{2t}{\beta})$ and some f' such that,

$$\begin{aligned} UB(\tilde{F}_p, f, x) &\leq LB(\tilde{F}_p, f', x) \quad \forall x, \\ F\big(x, f(x)\big) + t &\leq F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}(f'(x), \hat{f}(x)) - t, \\ F\big(x, f(x)\big) - F\big(x, \hat{f}(x)\big) + t &\leq \beta \cdot d^2_{\mathcal{Y}}(f'(x), \hat{f}(x)) - t, \\ \beta d^2_{\mathcal{Y}}(f(x), \hat{f}(x)) + t &\leq \beta \cdot d^2_{\mathcal{Y}}(f'(x), \hat{f}(x)) - t, \\ d^2_{\mathcal{Y}}(f'(x), \hat{f}(x)) &\geq 2t/\beta + d^2_{\mathcal{Y}}(f(x), \hat{f}(x)). \end{aligned}$$

Thus considering the greatest lower bound, any f' with $d^2_{\mathcal{Y}}(f'(x), \hat{f}(x)) \geq \frac{4t}{\beta}$ cannot be the minimizer of $\tilde{F}_p$, since there exists some other f with smaller distance from $\hat{f}$ that has smaller value compared to f'. The β dependence is expected, because if β is too small, i.e. F is sort of flat so the minimizer of $\tilde{F}$ might be far off from $\hat{f}$. □

B.2 When True Noise Distribution is not Available

Lemma 6. *Let* $\mathbf{Q}$, $\mathbf{P}$ *be the distributions defined in equation* (8)*, and $\tilde{d}_q(T, \tilde{Y})$ be the distance function as in equation 9, if* $\max_{ij} |\mathbf{P}_{ij} - \mathbf{Q}_{ij}| = \epsilon$ *then,*

$$\mathbb{E}_{\tilde{Y},\tilde{Z}\sim\mathbf{P}[:,y]}\big[\big|\tilde{d}_q(T,\tilde{Z}) - \tilde{d}_p(T,\tilde{Y})\big|\big] \leq \mathcal{O}\Big(k^2\Big(\sigma_{\max}(\mathbf{P}) + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big)\cdot\epsilon\Big) \qquad \forall y \in \mathcal{Y}. \tag{17}$$

Proof. Let $\tilde{\mathbf{d}}_q \in \mathbb{R}^k$ be a vector such that its i^{th} entry is given as $\tilde{\mathbf{d}}_q[i] = \tilde{d}_q(T, \tilde{Z} = y_i)$, and similarly, let $\mathbf{d}_p \in \mathbb{R}^k$ with $\mathbf{d}_p[i] = \tilde{d}_p(T, \tilde{Y} = y_i)$, and $\mathbf{d} \in \mathbb{R}^k$ with $\mathbf{d}[i] = d^2_{\mathcal{Y}}(T, Y = y_i)$. It is easy to see that, $\tilde{\mathbf{d}}_q = \mathbf{Q}^{-1}\mathbf{d}$ and $\tilde{\mathbf{d}}_p = \mathbf{P}^{-1}\mathbf{d}$. Now consider the following expectation w.r.t $\mathbf{P}$,

$$\mathbb{E}_{\mathbf{P}}[\tilde{\mathbf{d}}_q - \tilde{\mathbf{d}}_p] = \mathbb{E}_{\mathbf{P}}[\mathbf{Q}^{-1}\mathbf{d} - \mathbf{P}^{-1}\mathbf{d}] = \mathbf{P}\big(\mathbf{Q}^{-1}\mathbf{d} - \mathbf{P}^{-1}\mathbf{d}\big) = \mathbf{P}\big(\mathbf{Q}^{-1} - \mathbf{P}^{-1}\big)\mathbf{d}.$$

Let $\Delta\mathbf{P} = \mathbf{P} - \mathbf{Q}$, and using standard matrix inversion results for small perturbations, [Dem92], and $||\mathbf{d}||_\infty \leq 1$ we get

Since, $\max_{ij}(\Delta\mathbf{P})_{ij} \leq \epsilon$, we have $||\Delta\mathbf{P}||_2 \leq ||\Delta\mathbf{P}||_F \leq \epsilon k$

$$\begin{aligned}
||\mathbb{E}_{\mathbf{P}}[\tilde{\mathbf{d}}_p - \tilde{\mathbf{d}}_q]||_\infty &\leq ||\mathbf{P}||_2||(\mathbf{P} + \Delta\mathbf{P})^{-1} - \mathbf{P}^{-1}||_2||\mathbf{d}||_\infty, \\
&\leq ||\mathbf{P}||_2\Big(\kappa(\mathbf{P})||\mathbf{P}^{-1}||_2\frac{||\Delta\mathbf{P}||_2}{||\mathbf{P}||_2} + \mathcal{O}(||\Delta\mathbf{P}||_2^2)\Big), \\
&= \Big(\kappa(\mathbf{P})||\mathbf{P}^{-1}||_2||\Delta\mathbf{P}||_2\Big) + \mathcal{O}(||\Delta\mathbf{P}||_2^2), \\
&\leq \epsilon k \cdot \kappa(\mathbf{P})||\mathbf{P}^{-1}||_2 + \mathcal{O}(\epsilon^2 k^2), \\
&\leq \mathcal{O}\Big(k^2\Big(1 + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big) \cdot \epsilon\Big).
\end{aligned}$$

□

Lemma 7. *For $\tilde{F}_p$ and $\tilde{F}_q$ defined in* (10) *w.r.t. noise distributions $\mathbf{P}$ and $\mathbf{Q}$ respectively, and let $\max_{ij} |\mathbf{P}_{ij} - \mathbf{Q}_{ij}| \leq \epsilon$ then we have w.h.p.*

$$|\tilde{F}_p(x, y) - \tilde{F}_q(x, y)| \leq \tilde{\mathcal{O}}\Big(c_2\sqrt{\frac{1}{n}}\Big) + c_3\epsilon \qquad \forall y \in \mathcal{Y}. \tag{18}$$

with $c_2 = k^2\Big(1 + \frac{\kappa(\mathbf{P})}{\sigma^2_{\min}(\mathbf{P})}\Big)$ and $c_3 = k^2\Big(\sigma_{\max}(\mathbf{P}) + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big)$.

Proof. Recall the definitions,

$$\tilde{F}_p(x, y) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\tilde{d}_p(y, \tilde{y}_i), \qquad \tilde{F}_q(x, y) := \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\tilde{d}_q(y, \tilde{z}_i).$$

$$\tilde{F}_p(x, y) - \tilde{F}_q(x, y) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\Big(\tilde{d}_p(y, \tilde{y}_i) - \tilde{d}_q(y, \tilde{z}_i)\Big) = \frac{1}{n}\sum_{i=1}^{n} \alpha_i(x)\xi(y, \tilde{y}_i, \tilde{z}_i).$$

$\alpha_i(x) \cdot \xi(y, \tilde{y}_i, \tilde{z}_i)$ are independent r.v. and $\alpha_i(x) \leq 1$, but we don't know if the $\xi(y, \tilde{y}_i, \tilde{z}_i)$ are bounded. To see that $\xi(y, \tilde{y}_i, \tilde{z}_i)$ are bounded by $||\mathbf{Q}^{-1} - \mathbf{P}^{-1}||_2||\mathbf{d}||_\infty \leq c_2$ (see lemma 6) and from lemma 6, $\mathbb{E}[\xi(y, \tilde{y}_i, \tilde{z}_i)] \leq c_3\epsilon$, thus using Hoeffding's inequality gives the result. □

Lemma 8. *Let $\hat{f}_p$ be the minimizer as defined in equation 10 over the noisy labels drawn from $\mathbf{P}$, and let $\hat{f}_q$ (defined in eq. 10) be the minimizer over the noisy labels obtained from conditional distribution $\mathbf{Q}$ then w.h.p.*

$$d^2_{\mathcal{Y}}\big(\hat{f}_q(x), \hat{f}(x)\big) \leq \tilde{\mathcal{O}}\Big(\frac{1}{\beta}\big(c_1 + c2\big)\sqrt{\frac{1}{n}} + \frac{c_3}{\beta}\epsilon\Big) \qquad \forall x \in \mathcal{X}. \tag{19}$$

Proof. let $t_1 = O\Big(c_1\sqrt{\frac{1}{n}\log\Big(\frac{|\mathcal{Y}|}{\delta}\Big)}\Big)$ and $t_2 = \mathcal{O}\Big(c_2\sqrt{\frac{1}{n}\log\Big(\frac{|\mathcal{Y}|}{\delta}\Big)}\Big) + c_3\epsilon$, then combining lemma 7 and 4 we have,

$$F\big(x, f(x)\big) - t_1 - t_2 \leq \tilde{F}_q\big(x, f(x)\big) \leq F\big(x, f(x)\big) + t_1 + t_2.$$

Then following same argument as in lemma 5, we get the result. □

Theorem 2. *(Generalization Error) Let $\hat{f}$ be the minimizer as defined in* (2) *over the clean labels and let $\hat{f}_q$ (defined in* (10)*) be the minimizer over the noisy labels obtained from weak supervision inference in Algorithm 1. Suppose assumptions 2,3,4 hold. Then there exist constants $C_1, C_2 > 0$ dependent on $\sigma_{\max}(\mathbf{P}), \sigma_{\min}(\mathbf{P})$ and k such that w.h.p.,*

$$R(\hat{f}_q) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}) + \tilde{\mathcal{O}}\Big(\frac{C_1}{\beta} n^{-\frac{1}{2}}\Big) + \tilde{\mathcal{O}}\Big(\frac{C_2}{\beta}(\epsilon(d^+) + \epsilon(d^-))\Big). \tag{12}$$

Proof. Recall the definition of risk function,

$$R(f) = \mathbb{E}_{x,y}\big[d^2_{\mathcal{Y}}\big(f(x), y\big)\big].$$

$$\begin{aligned}
R(\hat{f}_q) &= \mathbb{E}_{x,y}\big[d^2_{\mathcal{Y}}\big(\hat{f}_q(x), y\big)\big], \\
&\leq \mathbb{E}_{x,y}\big[d^2_{\mathcal{Y}}\big(\hat{f}_q(x), \hat{f}(x)\big) + d^2_{\mathcal{Y}}(\hat{f}(x), y) + 2d_{\mathcal{Y}}(\hat{f}_q(x), \hat{f}(x)) \cdot d_{\mathcal{Y}}(\hat{f}(x), y)\big], \\
&= \mathbb{E}_x[d^2_{\mathcal{Y}}\big(\hat{f}_q(x), \hat{f}(x)\big)] + R(\hat{f}) + \tilde{\mathcal{O}}(n^{-1/4}), \\
&\leq \tilde{\mathcal{O}}\Big(\frac{1}{\beta}\big(c_1 + c2\big)\sqrt{\frac{1}{n}} + \frac{c_2}{\beta}\epsilon\Big) + R(\hat{f}) + \tilde{\mathcal{O}}(n^{-1/4}).
\end{aligned}$$

Using the result from [CRR16],

$$R(\hat{f}) \leq R(f^*) + \mathcal{O}(n^{-1/4}).$$

Combining the two we get

$$R(\hat{f}_q) \leq R(f^*) + \tilde{\mathcal{O}}(n^{-1/4}) + \tilde{\mathcal{O}}\Big(\frac{1}{\beta}\big(c_1 + c2\big)\sqrt{\frac{1}{n}} + \frac{c_3}{\beta}\epsilon)\Big).$$

We get the end result by plugging in the bound on $\epsilon = \max_{ij} ||\mathbf{P} - \mathbf{Q}||$ from lemma 11 and the bound on parameter recovery error $||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_\infty$ from Theorem 1.

□

Lemma 9. *The posterior distribution function $P_{\boldsymbol{\theta}}(Y = y|\Lambda = \Lambda^u)$ is $(2, \ell_\infty)-$Lipshcitz continuous in $\boldsymbol{\theta}$ for any $y \in \mathcal{Y}$ and $\Lambda^u \in \mathcal{Y}^m$.*

$$|P_{\boldsymbol{\theta}_1}(Y = y|\Lambda = \Lambda^u) - P_{\boldsymbol{\theta}_2}(Y = y|\Lambda = \Lambda^u)| \leq 2||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_\infty \qquad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^m.$$

Proof. Recall the definition of the posterior distribution,

$$P_{\boldsymbol{\theta}}(Y = y|\Lambda = \Lambda^u) = \frac{p(Y = y_i)P_{\boldsymbol{\theta}}(\Lambda = \Lambda^u|Y = y_i)}{\sum_{y_j \in \mathcal{Y}} p(Y = y_j)P_{\boldsymbol{\theta}}(\Lambda = \Lambda^u|Y = y_j)}.$$

For convenience let $\mathbf{d}^{(u,i)} \in \mathbb{R}^m$ be such that its a^{th} entry $\mathbf{d}^{(u,i)}_a = d^2_{\mathcal{Y}}(\Lambda^u_a, y_i)$

$$P_{\boldsymbol{\theta}}(Y = y|\Lambda = \Lambda^u) = \frac{P(Y = y_i)\exp(-\boldsymbol{\theta}^T\mathbf{d}^{(u,i)})}{\sum_{y_j \in \mathcal{Y}} P(Y = y_j)\exp(-\boldsymbol{\theta}^T\mathbf{d}^{(u,j)})}.$$

Let $Z_2(\boldsymbol{\theta}) = \sum_{y_j \in \mathcal{Y}} P(Y = y_j)\exp(-\boldsymbol{\theta}^T\mathbf{d}^{(u,j)})$, then

$$-\nabla_{\boldsymbol{\theta}} \log(Z_2(\boldsymbol{\theta})) = \frac{\sum_{y_j \in \mathcal{Y}} \mathbf{d}^{(u,j)} P(Y = y_j)\exp(-\boldsymbol{\theta}^T\mathbf{d}^{(u,j)})}{Z_2(\boldsymbol{\theta})} = \mathbb{E}_{Y|\Lambda}[\mathbf{d}].$$

Since distances are upper bounded by 1, $||\mathbf{d}||_\infty \leq 1$, so $||\mathbb{E}_{Y|\Lambda}[\mathbf{d}]||_\infty \leq 1$.
Now,

$$\nabla_{\boldsymbol{\theta}} \log\big(P_{\boldsymbol{\theta}}(Y = y|\Lambda = \Lambda^u)\big) = -\mathbf{d}^{(u,i)} - \nabla_{\boldsymbol{\theta}} \log(Z_2(\boldsymbol{\theta})).$$

Thus $||\nabla_{\boldsymbol{\theta}} \log \big(P_{\boldsymbol{\theta}}(Y = y|\Lambda = \Lambda^u)\big)||_\infty \leq 2$.

$$\implies |\log \big(P_{\boldsymbol{\theta}_1}(Y = y|\Lambda = \Lambda^u)\big) - \log \big(P_{\boldsymbol{\theta}_2}(Y = y|\Lambda = \Lambda^u)\big)| \leq 2||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_\infty.$$

Using the fact that for any $t_1, t_2 \in [0, 1]$ $|t_1 - t_2| \leq |\log(t_1) - \log(t_2)|$, gives us the result.

□

Lemma 10. *The distribution function $P_{\boldsymbol{\theta}}(\Lambda = \Lambda^u|Y = y)$ is $(2, \ell_\infty)-$Lipshcitz continuous in $\boldsymbol{\theta}$ for any $y \in \mathcal{Y}$ and $\Lambda^u \in \mathcal{Y}^m$.*

$$|P_{\boldsymbol{\theta}_1}(\Lambda = \Lambda^u|Y = y) - P_{\boldsymbol{\theta}_2}(\Lambda = \Lambda^u|Y = y)| \leq 2||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_\infty \qquad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^m.$$

Proof. Doing the same steps as in the proof of lemma 9 gives the result. □

Lemma 11. *For the noise distributions $\mathbf{P}, \mathbf{Q}$ in* (8) *with parameters $\boldsymbol{\theta}$, $\hat{\boldsymbol{\theta}}$ respectively and $\mathcal{Y}$ restricted only to the elements with non-zero prior probability, $\mathcal{Y}' = \{y \in \mathcal{Y} : P(Y = y) > 0\}$ the following holds,*

$$\max_{ij} |\mathbf{P}_{ij} - \mathbf{Q}_{ij}| \leq 4 \cdot k^m ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_\infty .$$

Proof. It is easy to see that for any two bounded functions f_1, f_2 with $|f_1(x)| \leq 1, |f_2(x)| \leq 1$ and Lipschitz continuous with constants L_1, L_2, the product of them is also Lipschitz continuous but with constant $L_1 + L_2$. Using this fact along with lemma 9 and lemma 10 gives the result,

$$|\mathbf{P}_{ij} - \mathbf{Q}_{ij}| \leq \sum_{\Lambda^u \in \mathcal{Y}'} |P_{\boldsymbol{\theta}}(y_i|\Lambda^u)P_{\boldsymbol{\theta}}(\Lambda^u|y_j) - P_{\hat{\boldsymbol{\theta}}}(y_i|\Lambda^u)P_{\hat{\boldsymbol{\theta}}}(\Lambda^u|y_j)| \leq 4 \cdot k^m ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}||_\infty.$$

□

It is important to note that we are restricting the values of y and λ to $\mathcal{Y}'$ which is the set of y with non-zero prior probability and by our assumption it is small.

C Proofs for Continuous Label Spaces

Next we present the proofs for the results in the continuous (manifold-valued) label spaces. We restate the first result on invariance:

Lemma 1. *For $\mathcal{Y} = \mathcal{M}$, a hyperbolic manifold, $y \sim P$ for some distribution P on $\mathcal{M}$ and labeling functions λ^a, λ^b drawn from* (3)*,*

$$\mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^a, \lambda^b) = \mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^b, y)\mathbb{E} \cosh d_{\mathcal{Y}}(\lambda^b, y),$$

while for $\mathcal{Y} = \mathcal{M}$ a spherical manifold,

$$\mathbb{E} \cos d_{\mathcal{Y}}(\lambda^a, \lambda^b) = \mathbb{E} \cos d_{\mathcal{Y}}(\lambda^b, y)\mathbb{E} \cos d_{\mathcal{Y}}(\lambda^b, y).$$

Proof. We start with the hyperbolic law of cosines, which states that

$$\cosh d(\lambda^a, \lambda^b) = \cosh d(\lambda^a, y) \cosh d(\lambda^b, y) + \sinh d(\lambda^a, y) \sinh d(\lambda^b, y) \cos \alpha,$$

where α is the angle between the sides of the triangle formed by (y, λ^a) and $(y, \lambda^b$. We can rewrite this as follows. Let $v^a = \log_y(\lambda^a)$, $v^b = \log_y(\lambda^b$ be tangent vectors in $T_y M$. Then,

$$\cosh d(\lambda^a, \lambda^b) = \cosh d(\lambda^a, y) \cosh d(\lambda^b, y) + (\sinh \|v^a\| \sinh \|v^b\|)\langle \frac{v^a}{\|v^a\|}, \frac{v^b}{\|v^b\|} \rangle.$$

Next, we take the expectation conditioned on y. The right-most term is then

$$\begin{aligned}
&\mathbb{E}[(\sinh\|v^a\|\sinh\|v^b\|)\langle\frac{v^a}{\|v^a\|},\frac{v^b}{\|v^b\|}\rangle|y] \\
&\quad=\mathbb{E}[(\sinh\|v^a\|\sinh\|v^b\|)|y]\mathbb{E}[\langle\frac{v^a}{\|v^a\|},\frac{v^b}{\|v^b\|}\rangle|y] \\
&\quad=0,
\end{aligned}$$

where the last equality follows from the fact that v^a and v^b are independent conditioned on y. This leaves us with the $\cosh$ product terms. Taking expectation again with respect to y gives the result.

The spherical version of the result is nearly identical, replacing hyperbolic sines and cosines with sines and cosines, respectively. □

Note, in addition, that it is easy to obtain a version of this result for curvatures that are not equal to -1 in the hyperbolic case (or $+1$ in the spherical case).

We will use this result for our consistency result, restated below.

Theorem 3. *Let $\mathcal{M}$ be a hyperbolic manifold. Fix $0<\delta<1$ and let $\Delta(\delta)=\min_\rho Pr(\forall i, d_{\mathcal{Y}}(\lambda^a(i),\lambda^b(i)\leq\rho))\geq 1-\delta$. Then, there exists a constant C_1 so that with probability at least $1-\delta$,*

$$\mathbb{E}|\hat{\mathbb{E}}d^2_{\mathcal{Y}}(\lambda^a,y))-\mathbb{E}d^2_{\mathcal{Y}}(\lambda^a,y)|\leq\frac{C_1\cosh(\Delta(\delta))^{3/2}}{C_0\sqrt{2n}}.$$

Proof. First, we will condition on the event that the observed outputs have maximal distance (i.e., diameter) Δ. This implies that our statements hold with high probability. Then, we use McDiarmid's inequality. For each pair of distinct LFs a,b, we have that

$$P\left(\frac{1}{n}|\sum_{i=1}^n\cosh(d(\lambda^a(i),\lambda^b(i)))-\mathbb{E}\cosh(d(\lambda^a,\lambda^b))|\geq t\right)\leq 2\exp\left(-\frac{2nt^2}{\cosh(\Delta)}\right),$$

Integrating the expression above in t, we obtain

$$\mathbb{E}|\hat{\mathbb{E}}\cosh(d(\lambda^a,\lambda^b))-\mathbb{E}\cosh(d(\lambda^a,\lambda^b))|\leq\frac{\sqrt{\pi\cosh(\Delta)}}{\sqrt{2n}}. \tag{20}$$

Next, we use this to control the gap on our estimator. Recall that using the triplet approach, we estimate

$$\hat{\mathbb{E}}\cosh(d(\lambda^a,y))=\sqrt{\frac{\hat{\mathbb{E}}\cosh d(\lambda^a,\lambda^b)\hat{\mathbb{E}}\cosh d(\lambda^a,\lambda^c)}{(\hat{\mathbb{E}}d(\lambda^b,\lambda^c))^2}}.$$

For notational convenience, we write $\nu(a)$ for $\mathbb{E}(\cosh(d(\lambda^a,y)))$, $\hat{\nu}(a)$ for its empirical counterpart, and $\nu(a,b)$ and $\hat{\nu}(a,b)$ for the versions between pairs of LFs a,b. Then, the above becomes

$$\hat{\nu}(a)=\sqrt{\frac{\hat{\nu}(a,b)\hat{\nu}(a,c)}{(\hat{\nu}(b,c))^2}}.$$

Note that $\cosh(x)\geq 1$, so that $\hat{\nu}(a,b)\geq 1$ and similarly for the empirical versions. We also have that $\hat{\nu}(a,b)\leq\cosh(\Delta)$. With this, we can begin our perturbation analysis. Applying Lemma 1, we

have that

$$\begin{aligned}
\mathbb{E}|\hat{\nu}(a)-\nu(a)| &= E\left|\sqrt{\frac{\hat{\nu}(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&= \mathbb{E}\left|\sqrt{\frac{\hat{\nu}(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}+\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&\leq \mathbb{E}\left|\sqrt{\frac{\hat{\nu}(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}\right|+\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&= \mathbb{E}\left|\sqrt{\frac{\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}(\sqrt{\hat{\nu}(a,b)}-\sqrt{\nu(a,b)})\right|+\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&\leq \frac{\sqrt{\pi}\cosh(\Delta^2)}{\sqrt{2n}}+\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right|.
\end{aligned}$$

To see why the last step holds, note that $\sqrt{\hat{\nu}(a,c)} \leq \sqrt{\cosh(\Delta)}$, while $\hat{\nu}(b,c) \geq 1$. Next, for $\alpha, \beta \geq 1$, $\sqrt{\alpha}-\sqrt{\beta} = \frac{\alpha-\beta}{\sqrt{\alpha}-\sqrt{\beta}} \leq \alpha-\beta$. This means that $\mathbb{E}|\sqrt{\hat{\nu}(a,b)}-\sqrt{\nu(a,b)}| \leq \mathbb{E}|\hat{\nu}(a,b)-\nu(a,b)| \leq \frac{\sqrt{\pi\cosh(\Delta)}}{\sqrt{2n}}$ using (20).

Now we can continue, adding and subtracting as before. We have that

$$\begin{aligned}
&\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&\quad\leq \mathbb{E}\left|\sqrt{\frac{\nu(a,b)\hat{\nu}(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\hat{\nu}(b,c)^2}}\right|+\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\nu(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&\quad\leq \frac{\sqrt{\pi}\cosh(\Delta)}{\sqrt{2n}}+\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\nu(a,c)}{\hat{\nu}(b,c)^2}}-\sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \\
&\quad\leq \frac{\sqrt{\pi}\cosh(\Delta)}{\sqrt{2n}}+\frac{\sqrt{\pi}\cosh(\Delta)^{3/2}}{\sqrt{2n}}.
\end{aligned}$$

Putting it all together, with probability at least $1-\delta$,

$$\mathbb{E}|\hat{\mathbb{E}}\cosh(d(\lambda^a,y))-\mathbb{E}\cosh(d(\lambda^a,y))| \leq \frac{2\sqrt{\pi}\cosh(\Delta)+\sqrt{\pi}\cosh(\Delta)^{3/2}}{\sqrt{2n}}. \tag{21}$$

Next, recall that C_0 satisfies $\mathbb{E}|\hat{\mathbb{E}}\cosh(d(\lambda^a,\lambda^b))-\mathbb{E}\cosh(d(\lambda^a,\lambda^b))| \geq C_0\mathbb{E}|\hat{\mathbb{E}}d(\lambda^a,\lambda^b))-\mathbb{E}d(\lambda^a,\lambda^b)|$. Thus,

$$\mathbb{E}|\hat{\mathbb{E}}d^2(\lambda^a,y)-\mathbb{E}d^2(\lambda^a,y)| \leq \frac{2\sqrt{\pi}\cosh(\Delta)+\sqrt{\pi}\cosh(\Delta)^{3/2}}{C_0\sqrt{2n}}.$$

This concludes the proof. □

Next, we will prove a simple result that is needed in the proof of Theorem 5. Consider the distribution P of the quantities $\alpha(x)(y)d^2_{\mathcal{Y}}(z,y)$ for some fixed $z \in \mathcal{M}$. We can think of this as the population-level version of sample distances that are observed in the supervised version of the problem. We do not have access to it in our approach; it will be used only as an object in our proof. Recall we set $q = \arg\min_{z\in\mathcal{Y}} \mathbb{E}[\alpha(x)(y)d^2_{\mathcal{Y}}(z,y)]$ to be the population-level minimizer. Here we use the notation $\alpha(x)(y)$ to denote the corresponding kernel value at a point y. Finally, let us denote P' to be the distribution over the quantities $\alpha(x)(y)\sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\lambda_{a,i})$.

Lemma 12. *Let the distributions P and P' be defined as above, with q the minimizer of $\mathbb{E}_P[\alpha(x)(y)d^2_{\mathcal{Y}}(z,y)]$. Suppose that Assumptions 5 and 6 hold. Then, q is also the minimizer of $\mathbb{E}'_P[\alpha(x)(y)\sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\lambda_{a,i})]$.*

Proof. We will use a simple symmetry argument. First, note that we can write q in the following way,

$$q = \arg\min_{z\in\mathcal{Y}} \int_{T_q\mathcal{M}} \alpha(x)(\log_q(v))d^2_{\mathcal{Y}}(z,\exp_q(v))dP.$$

Since $\mathcal{M}$ is a symmetric manifold, if $v \in T_q\mathcal{M}$, there is an isometry sending v to $-v \in T_q\mathcal{M}$. Using this isometry and Assumption 6, we can also write

$$q = \arg\min_{z\in\mathcal{Y}} \int_{T_q\mathcal{M}} \alpha(x)(\log_q(-v))d^2_{\mathcal{Y}}(z,\exp_q(-v))dP.$$

Our approach will be to formulate similar symmetric expressions for the minimizer, but this time for the loss over the distribution P'. We will then be able to show, using triangle inequality, that q remains the minimizer.

We can similarly express the minimizer of the loss for P' as

$$\arg\min_{z\in\mathcal{Y}} \int_{T_q\mathcal{M}} \int_{(T_{\exp_q(v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\exp_{\exp_q(v)}(v^a))dP'.$$

Here we have broken down the expectation over P' by applying the tower law; the inner expectation is conditioned on point $\exp_q(v)$ and runs over the labeling function outputs $\lambda^1,\ldots,\lambda^m$.

Again using Assumption 6, we can write the minimizer for the loss over P' as $\arg\min_{z\in\mathcal{Y}} F'(z)$, where

$$F'(z) = \int_{T_q\mathcal{M}} \int_{(T_{\exp_q(-v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(-v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\exp_{\exp_q(-v)}(-v^a))dP'.$$

Thus we can also write the minimizer as $\arg\min_{z\in\mathcal{Y}} F'(z)$, where

$$F'(z) = \int_{T_q\mathcal{M}} \int_{(T_{\exp_q(-v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(-v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\exp_{\exp_q(-v)}(-v^a))dP'.$$

With this, we can write

$$\begin{aligned}
F'(z) = \frac{1}{2}\Bigg(&\int_{T_q\mathcal{M}} \int_{(T_{\exp_q(v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\exp_{\exp_q(v)}(v^a))dP' \\
&+ \int_{T_q\mathcal{M}} \int_{(T_{\exp_q(-v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(-v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}(z,\exp_{\exp_q(-v)}(-v^a))dP'\Bigg) \\
= \frac{1}{2}\Bigg(&\int_{T_q\mathcal{M}} \int_{(T_{\exp_q(v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(v)) \sum_{a=1}^m \beta_a^2 d^2_{\mathcal{Y}}\Big((z,\exp_{\exp_q(v)}(v^a)) \\
&+ d^2_{\mathcal{Y}}(z,\exp_{\exp_q(-v)}(PT_{\exp_q(v)\to\exp_q(-v)}(-v^a)))\Big)dP'\Bigg),
\end{aligned}$$

where $PT_{p\to s}$ denotes parallel transport from p to s.

Note that q is on the geodesic between $\exp_{\exp_q(v)}(v^a)$ and $\exp_{\exp_q(-v)}(PT_{\exp_q(v)\to\exp_q(-v)}(-v^a))$. We exploit this fact by applying the following squared-distance inequality. For three points p, s, z, from the triangle inequality,

$$d_{\mathcal{Y}}(p,z) + d_{\mathcal{Y}}(s,z) \geq d_{\mathcal{Y}}(p,s).$$

Squaring both sides and applying

$$d^2_{\mathcal{Y}}(p,z) + d^2_{\mathcal{Y}}(s,z) \geq 2d_{\mathcal{Y}}(p,z)d_{\mathcal{Y}}(s,z),$$

we obtain that

$$2(d^2_{\mathcal{Y}}(p,z) + d^2_{\mathcal{Y}}(s,z)) \geq d^2_{\mathcal{Y}}(p,s),$$

so that

$$d^2_{\mathcal{Y}}(p,z) + d^2_{\mathcal{Y}}(q,z) \geq \frac{1}{2} d^2_{\mathcal{Y}}(p,q).$$

Setting p to be $\exp_{\exp_q(v)}(v^a)$ and s to be $\exp_{\exp_q(-v)}(PT_{\exp_q(v)\to\exp_q(-v)}(-v^a))$ in the above gives

$$F'(z) \geq \frac{1}{2}\left(\int_{T_q\mathcal{M}} \int_{(T_{\exp_q(v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(v)) \sum_{a=1}^{m} \beta_a^2 \right.$$
$$\left. \frac{1}{2} d^2_{\mathcal{Y}}(\exp_{\exp_q(v)}(v^a), \exp_{\exp_q(-v)}(PT_{\exp_q(v)\to\exp_q(-v)}(-v^a)))dP'\right).$$

Now we can apply the fact that q is on the geodesic to rewrite this as

$$F'(z) \geq \frac{1}{2}\left(\int_{T_q\mathcal{M}} \int_{(T_{\exp_q(v)}\mathcal{M})^{\otimes m})} \alpha(x)(\log_q(v)) \sum_{a=1}^{m} \beta_a^2 \frac{1}{2} 4 d^2_{\mathcal{Y}}(q, \exp_{\exp_q(v)}(v^a))dP'\right).$$

This is because the length of the geodesic connecting $\exp_{\exp_q(v)}(v^a)$ and $\exp_{\exp_q(-v)}(PT_{\exp_q(v)\to\exp_q(-v)}(-v^a))$ is twice that of the geodesic connecting $\exp_{\exp_q(v)}(v^a)$ to q.

Thus, we have

$$F'(z) \geq F'(q),$$

and we are done. □

Finally, this enables us to prove our main result, Theorem 5, restated below:

Theorem 5. *Let $\mathcal{M}$ be a complete manifold and suppose the assumptions above hold. Then, there exist constants C_3, C_4*

$$\mathbb{E}[d^2_{\mathcal{Y}}(\hat{f}(x), \tilde{f}(x))] \leq \frac{C_3\sigma_o^2}{nk_{\min}} + \frac{C_4 \sum_{a=1}^{m} \beta_a^2 \hat{\mu}_a^2}{mnk_{\min}}.$$

Proof. We use Lemma 12 and compute a bound on the expected distance from the empirical estimates to the common center. In both cases, the approach is nearly identical to that of [Str20] (proof of Theorem 3.2.1); we include these steps for clarity. Suppose that the minimum and maximum values of α are $\alpha_{\min}$ and $\alpha_{\max}$, respectively.

Then, letting we have that, using the hugging function assumption

$$\|\log_q(\hat{f}(x)) - \log_q(y_i)\|^2 \leq k_{\min} d^2_{\mathcal{Y}}(q, \hat{f}(x)) + d^2_{\mathcal{Y}}(\hat{f}(x), y_i).$$

We also have that

$$\|\log_q(\hat{f}(x)) - \log_q(y_i)\|^2 = d^2_{\mathcal{Y}}(q, \hat{f}(x)) - 2\langle \log_q(\hat{f}(x)), \log_q(y_i)\rangle + d^2_{\mathcal{Y}}(q, y_i).$$

Then,

$$(1 - k_{\min}) d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 2\langle \log_q(\hat{f}(x)), \log_q(y_i)\rangle + d^2_{\mathcal{Y}}(\hat{f}(x), y_i) - d^2_{\mathcal{Y}}(q, y_i).$$

Now, multiply each of the equations by α_i and sum over them. In that case, the different on the right side is non-positive, as $\hat{f}(x)$ is the empirical minimizer. This yields

$$\sum_{i=1}^{n} \alpha(x)_i (1 - k_{\min}) d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq \sum_{i=1}^{n} \alpha(x)_i 2\langle \log_q(\hat{f}(x)), \log_q(y_i)\rangle.$$

Using the minimum and maximum values of α, and setting $\bar{q} = \sum_{i=1}^{n} \log_q(y_i)$, we get

$$\alpha_{\min}(1 - k_{\min})d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 2\alpha_{\max}\langle \log_q(\hat{f}(x)), \bar{q}\rangle.$$

We can apply Cauchy-Schwarz, simplify, then square, obtaining

$$\alpha^2_{\min}(1 - k_{\min})^2 d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 4\alpha^2_{\max}\|\bar{q}\|^2.$$

What remains is to take expectation and use the fact that the tangent vectors summed up to form $\bar{q}$ are independent. This yields

$$\alpha^2_{\min}(1 - k_{\min})^2 \mathbb{E}d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 4\alpha^2_{\max}\frac{\sigma_o^2}{n}.$$

Thus we obtain

$$\alpha^2_{\min}(1 - k_{\min})^2 \mathbb{E}d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 4\alpha^2_{\max}\frac{\sigma_o^2}{n},$$

or

$$\mathbb{E}d^2_{\mathcal{Y}}(q, \hat{f}(x)) \leq 4\frac{\alpha^2_{\max}}{\alpha^2_{\min}}\frac{\sigma_o^2}{nk_{\min}}. \tag{22}$$

We use the same approach, but this apply it to the $m \times n$ points given by the LFs drawn from distribution P'. This yields

$$\alpha^2_{\min}(1 - k_{\min})^2 \mathbb{E}d^2_{\mathcal{Y}}(q, \tilde{f}(x)) \leq 4\alpha^2_{\max}\frac{\sum_{i=1}^{m}\beta_a^2\sigma_a^2}{mn},$$

where σ_a^2 corresponds to the expected squared distance for LF a to q. We bound this with triangle inequality, obtaining $\sigma_a^2 \leq 2\sigma_o^2 + 2\hat{\mu}_a^2$, so that

$$\alpha^2_{\min}(1 - k_{\min})^2 \mathbb{E}d^2_{\mathcal{Y}}(q, \tilde{f}(x)) \leq 8\alpha^2_{\max}\frac{\sum_{i=1}^{m}\beta_a^2(\sigma_o + \hat{\mu}_a^2}{mn},$$

or,

$$\mathbb{E}d^2_{\mathcal{Y}}(q, \tilde{f}(x)) \leq 8\frac{\alpha^2_{\max}}{\alpha^2_{\min}}\frac{\sum_{i=1}^{m}\beta_a^2(\sigma_o + \hat{\mu}_a^2}{mnk_{\min}}. \tag{23}$$

Now, again using triangle inequality,

$$\mathbb{E}d^2_{\mathcal{Y}}(\hat{f}(x), \tilde{f}(x)) \leq 2\mathbb{E}d^2_{\mathcal{Y}}(q, \hat{f}(x)) + 2\mathbb{E}d^2_{\mathcal{Y}}(q, \tilde{f}(x)).$$

Plugging (23) and (22) into this bound produces the result. □

D Additional Continuous Label Space Details

We provide some additional details on the continuous (manifold-valued) case.

Computing $\Delta(\delta)$ In Theorem 3, we stated the result in terms of $\Delta(\delta)$, a quantity that trades off the probability of failure δ for the diameter of the largest ball that contains the observed points. Note that if we fix the curvature of the manifold, it is possible to compute an exact bound for this quantity by using formulas for the sizes of balls in d-dimensional manifolds of fixed curvature.

Hugging number Note that it is possible to derive a lower bound on the hugging number as a function of the curvature. The way to do so is to use *comparison theorems* that upper bound triangle edge lengths with those of larger-curvature triangles. This makes it possible to establish a concrete value for $k_{\min}$ as a function of the curvature.

We note, as well, that an upper bound $k_{\max}$ on the hugging number can be obtained by a simple rearrangement of Lemma 6 from [ZS16]. This result follows from a curvature lower bound based on hyperbolic law of cosines; the bound we describe follows from the opposite—an upper bound based on spherical triangles.

β Weights and Suboptimality An intuitive way to think of the estimator we described is the following simple Euclidean version. Suppose we have labeling functions $\lambda_1, \ldots, \lambda_m$ that are equal to $y + \varepsilon_a$, where $\varepsilon_a \sim \mathcal{N}(0, \sigma_a^2)$. In this case, if we seek an unbiased estimator with lowest variance, we require a set of weights β_a so that $\sum_a \beta_a = 1$ and $\text{Var}[\frac{1}{m}\sum_{a=1}^m \beta_a \lambda_a]$ is minimized. It is not hard to derive a closed-form solution for the β_a coefficients as a function of the terms σ_a^2.

Now, suppose we use the same solution, but with noisy estimates $\hat{\sigma}^2$ instead. Our weights $\hat{\beta}$ will yield a suboptimal variance, but this will not affect the scaling of the rate in terms of the number of samples n.

E Extended Background on Pseudo-Euclidean Embeddings

Finally, we provide some additional background on pseudo-metric spaces and pseudo-Euclidean embedding.

E.1 Pseudo-metric Spaces

Pseudo-metric spaces generalize metric spaces by removing the requirement that pairs of points at distance zero must be identical:

Definition 1. ***(Pseudo-metric Space)*** *A set $\mathcal{Y}$ along with a distance function $d_{\mathcal{Y}} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is called pseudo-metric space if $d_{\mathcal{Y}}$ satisfies the following conditions,*

$$\forall \mathbf{y}, \mathbf{z} \in \mathcal{Y} \qquad d_{\mathcal{Y}}(\mathbf{y}, \mathbf{z}) = d_{\mathcal{Y}}(\mathbf{y}, \mathbf{z}) \tag{24}$$

(Symmetry)

$$\forall \mathbf{y} \in \mathcal{Y} \qquad d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}) = 0 \tag{25}$$

(Reflexivity)

A finite pseudo-metric space has $|\mathcal{Y}| < \infty$.

E.2 Pseudo-Euclidean Spaces

The following definitions are for *finite-dimensional* vector spaces defined over the field $\mathbb{R}$.

Definition 2. ***(Symmetric Bilinear Form / Generalized Inner Product)*** *For a vector space $\mathcal{Y}$ over the field $\mathbb{R}$, a symmetric bilinear form is a function $\phi : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}$ satisfying the following properties $\forall y_1, y_2, z, y \in \mathcal{Y}, c \in \mathbb{R}$:*

P1) $\phi(y_1 + y_2, y) = \phi(y_1, y) + \phi(y_2, y)$,

P2) $\phi(cy, z) = c\phi(y, z)$,

P3) $\phi(y, z) = \phi(z, y)$.

Definition 3. ***(Squared Distance w.r.t. ϕ)*** *Let V be a real vector space equipped with generalized inner product ϕ, then the squared distance w.r.t. ϕ between any two vectors $\mathbf{y}, \mathbf{z} \in V$ is defined as,*

$$||\mathbf{y} - \mathbf{z}||_\phi^2 := \phi(\mathbf{y} - \mathbf{z}, \mathbf{y} - \mathbf{z})$$

This definition also gives a notion of squared length for every $\mathbf{y} \in V$,

$$||\mathbf{y}||_\phi^2 := \phi(\mathbf{y}, \mathbf{y})$$

The inner product can also be expressed in terms of a basis of the vector space V. Let the dimension of $\mathcal{Y}$ be d, and $\{\mathbf{b}_i\}_{i=1}^d$ be a basis of $\mathcal{Y}$, then for any two vectors $\mathbf{y} = [y_1, \ldots y_d], \mathbf{z} = [z_1, \ldots z_d] \in V$,

$$\phi(\mathbf{y}, \mathbf{z}) = \sum_{i=1}^d \sum_{j=1}^d y_i z_i \phi(\mathbf{b}_i, \mathbf{b}_j)$$

The matrix $\mathbf{M}(\phi) := [\phi(\mathbf{b}_i, \mathbf{b}_j)]_{1 \leq i,j \leq d}$ is called *the matrix of* ϕ w.r.t the basis $\{\mathbf{b}_i\}_{i=1}^d$ It gives a convenient way to express the inner product as $\phi(\mathbf{y}, \mathbf{z}) = \mathbf{y}^T \mathbf{M}(\phi)\mathbf{z}$. A symmetric bilinear form ϕ on a vector space of dimension d, is said to be *non-degenerate* if the rank of $\mathbf{M}(\phi)$ w.r.t to some basis is equal to d.

Example: For the $d-$ dimensional euclidean space with standard basis and ϕ as dot product we get $\mathbf{M}(\phi) = \mathbf{I}_d$

Definition 4. ***(Pseudo-euclidean Spaces)*** *A real vector space $\mathbb{R}^{d^+,d^-}$ of dimension $d = d^+ + d^-$, equipped with a non-degenerate symmetric bilinear form ϕ is called a pseudo-euclidean (or Minkowski) vector space of signature (d^+, d^-) if the matrix of ϕ w.r.t a basis $\{\mathbf{b}_i\}_{i=1}^d$ of $\mathbb{R}^{d^+,d^-}$, is given as,*

$$\mathbf{M}(\phi) = \begin{pmatrix} \mathbf{I}_{d^+} & \mathbf{0} \\ \mathbf{0} & -\mathbf{I}_{d^-} \end{pmatrix}_{d \times d}$$

Lastly, the tool that we used to ensure we have access to isometric embeddings is

Proposition 1. *([Gol85]) Let $\mathcal{Y} = \{y_0, \dots y_k\}$ be a finite pseudo-metric space equipped with distance function $d_{\mathcal{Y}}$, and let $\mathbf{V} = \{\mathbf{v}_i, \dots, \mathbf{v}_k\}$ be a collection of vectors in $\mathbb{R}^{d^+,d^-}$. Then $\mathcal{Y}$ is isometrically embedable in $\mathbb{R}^{d^+,d^-}$ if and only if,*

$$\langle \mathbf{v}_i, \mathbf{v}_j \rangle_\phi = \frac{1}{2}\Big(d^2_{\mathcal{Y}}(y_i, y_0) + d^2_{\mathcal{Y}}(y_j, y_0) - d^2_{\mathcal{Y}}(y_i, y_j)\Big) \quad \forall i, j \in [k] \tag{26}$$