Lifting Weak Supervision To Structured Prediction

Harit Vishwakarma
hvishwakarma@wisc.edu

Frederic Sala
fredsala@cs.wisc.edu

Department of Computer Sciences,
University of Wisconsin-Madison, WI, USA.

Abstract

Weak supervision (WS) is a rich set of techniques that produce pseudolabels by aggregating easily obtained but potentially noisy label estimates from a variety of 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. Empirical evaluation validates our claims and shows the merits of the proposed method[1].

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.

[1] https://github.com/SprocketLab/WS-Struct-Pred

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

An important advantage of WS in the standard setting of binary classification is that it sometimes 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 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,
- Experiments confirming the theoretical results and showing improvements over [SLV$^+$22].

2 Background and Problem Setup

Our goal is to theoretically characterize how well learning with pseudolabels (built with weak supervision techniques) performs in structured prediction. We seek to understand the interplay between the noise in WS sources and the generalization performance of the downstream structured prediction model. We provide brief background and 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, including $\mathcal{Y} = \mathbb{S}_d$, the sphere, or $\mathbb{H}_d$, the hyperboloid.

Learning and Generalization in Structured Prediction In conventional supervised learning we have a dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ of i.i.d. samples drawn from distribution ρ over $\mathcal{X} \times \mathcal{Y}$. 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 $\hat{f}$ is consistent:

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

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. The procedure here is to first compute the weights α and then to perform the optimization in (2) to make a prediction.

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, as long as there is no label noise. The key question we tackle is *does the use of pseudolabels instead of true labels y_i affect the generalization rate?*

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 $\lambda_{1,i}, \ldots, \lambda_{m,i}$. 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:

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

Z is the normalizing partition function, $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_m]^T > 0$ are *canonical* parameters, and E is a set of correlations. The model can be described in terms of the *mean* parameters $\mathbb{E}[d^2_{\mathcal{Y}}(\lambda_a, y)]$. Intuitively, if θ_a is large, the typical distance from λ_a to y is small and the LF is reliable; if θ_a is small, the LF is unreliable. This model is appropriate for several reasons. It is an exponential family model with useful theoretical properties. It subsumes popular special cases of noise, including, for regression, zero-mean multivariate Gaussian noise; for permutations, a generalization of the popular Mallows model; for the binary case, it produces a close relative of the Ising model.

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 $\hat{\theta}_a$ are used to weight the labeling functions, ensuring that more reliable votes receive a larger weight.

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

Do there exist estimation approaches yielding $\hat{\boldsymbol{\theta}}$ 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 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.

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. 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.

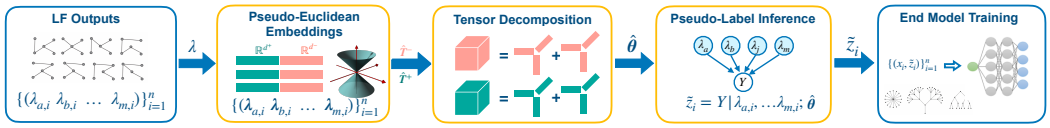

Figure 1: Illustration of our weak supervision pipeline for the finite label space setting.

The intuition behind this approach is the following. First, we need a technique that can provide consistent or nearly-consistent estimates of the parameters in the noise model. Second, we need to handle any finite metric space. Techniques like the one introduced in [FCS+20] handle the first—but do not work for generic finite metric spaces, only binary labels and certain sequences. Techniques like the one in [SLV+22] handle any metric space—but only have consistency guarantees in highly restrictive settings (e.g., it requires an isometric embedding, that the distribution over the resulting embeddings is isomorphic to certain distributions, the true label only takes on two values). Pseudo-Euclidean embeddings used with tensor decomposition algorithms meet both requirements

3.1 Pseudo-Euclidean Embeddings

Our first task is to embed the metric space into a continuous space—enabling easier computation and potential dimensionality reduction. A standard approach is multi-dimensional scaling (MDS) [KW78], which embeds $\mathcal{Y}$ into $\mathbb{R}^d$. A downside of MDS is that not all metric spaces embed (isometrically) into Euclidean space, as the square distance matrix $\mathbf{D}$ must be 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. Such spaces have been applied to similarity-based learning methods [PPD01, LRBM06, PHD+06]. 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 $\mathbf{u}, \mathbf{v}$ are $\langle \mathbf{u}, \mathbf{v} \rangle_\phi = \langle \mathbf{u}^+, \mathbf{v}^+ \rangle - \langle \mathbf{u}^-, \mathbf{v}^- \rangle$ and $d^2_\phi(\mathbf{u}, \mathbf{v}) = ||\mathbf{u}^+ - \mathbf{v}^+||_2^2 - ||\mathbf{u}^- - \mathbf{v}^-||_2^2$. These properties enable isometric embeddings: the distance can be decomposed into two components that are individually induced from p.s.d. inner products—and can thus be embedded via MDS. Indeed, pseudo-Euclidean embeddings effectively run MDS for each component (see Algorithm 1 steps 4-9). To recover the original distance, we obtain $||\mathbf{u}^+ - \mathbf{v}^+||_2^2$ and $||\mathbf{u}^- - \mathbf{v}^-||_2^2$ and subtract.

Example: To see why such embeddings are advantageous, we compare with a one-hot vector representation (whose dimension is $|\mathcal{Y}|$). 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. The pseudo-Euclidean embedding dimension is just $d = 3$; see Appendix for more details. 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, respectively. 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 $\boldsymbol{\Sigma}_{a,y}$. Our goal is to obtain an accurate estimate $\hat{\boldsymbol{\mu}}_{a,y} = \hat{\mathbb{E}}[\boldsymbol{\lambda}_a | \mathbf{y}]$, which we will use to estimate the mean parameters $\mathbb{E}[d^2_{\mathcal{Y}}(\lambda_a, y)]$. If we could observe y, it would be easy to empirically estimate $\boldsymbol{\mu}_{a,y}$—but we do not have access to it. Our approach will be to apply tensor decomposition for multi-view mixtures [AGJ14].

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 whose use is described in Assumption 3.

Algorithm 1 Algorithm for Pseudolabel Construction

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

▷ Step 1: Compute pseudo-Euclidean Embeddings
Construct matrices $\mathbf{D} \in \mathbb{R}^{r\times r}$, $\mathbf{D}_{ij} = d^2_{\mathcal{Y}}(y_i, y_j)$ and $\mathbf{M} \in \mathbb{R}^{r\times r}$, $\mathbf{M}_{ij} = \frac{1}{2}(\mathbf{D}^2_{0i} + \mathbf{D}^2_{0j} - \mathbf{D}^2_{ij})$
Compute eigendecomposition of $\mathbf{M}$ and let $\mathbf{M} = \mathbf{U}\mathbf{C}\mathbf{U}^T$
Set l^+, l^- be indices of positive and negative eigenvalues sorted by their magnitude
Let $d^+ = |l^+|, \quad d^- = |l^-|$ i.e. the sizes of lists l^+ and l^- respectively.
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 $\boldsymbol{\lambda}_{a,i} = g(\lambda_{a,i}), \boldsymbol{\lambda}_{b,i} = g(\lambda_{b,i}), \boldsymbol{\lambda}_{c,i} = g(\lambda_{c,i}) \quad \forall i \in [n]$ where a, b, c are uncorrelated
    Construct tensors $\hat{\mathbf{T}}^+$ and $\hat{\mathbf{T}}^-$ as defined in (5) for triplet (a, b, c)
    $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^+_{b,y}, \hat{\boldsymbol{\mu}}^+_{c,y}$ = `TensorDecomposition`$(\hat{\mathbf{T}}^+)$
    $\hat{\boldsymbol{\mu}}^-_{a,y}, \hat{\boldsymbol{\mu}}^-_{b,y}, \hat{\boldsymbol{\mu}}^-_{c,y}$ = `TensorDecomposition`$(\hat{\mathbf{T}}^-)$
    $s^+_{a,y} = \min_{z\in\{-1,+1\}} \phi(z \cdot \hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+)$ and similarly $s^+_{b,y}, s^+_{c,y}, s^-_{a,y}, s^-_{b,y}, s^-_{c,y}$
    $\hat{\boldsymbol{\mu}}^+_{a,y} = s^+_{a,y} \cdot \hat{\boldsymbol{\mu}}^+_{a,y}$ and similarly correct signs of $\hat{\boldsymbol{\mu}}^+_{b,y}, \hat{\boldsymbol{\mu}}^+_{c,y}, \hat{\boldsymbol{\mu}}^-_{a,y}, \hat{\boldsymbol{\mu}}^-_{b,y}, \hat{\boldsymbol{\mu}}^-_{c,y}$
end for

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

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 such as tensors of outer products of LF triplets. We require that for each a such a triplet exists. Then,

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

Here w_y are the mixture probabilities (prior probabilities of Y) and $\mathcal{Y}_s = \{y : w_y > 0\}$. We similarly define $\mathbf{T}^-$ and $\hat{\mathbf{T}}^-$. We then obtain estimates $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ using an algorithm from [AGH+14] with minor modifications to handle pseudo-Euclidean rather than Euclidean space. The overall approach is shown in Algorithm 1. We have three key assumptions for our analysis,

Assumption 1. *The support of P_Y, i.e., $k = |\{y : w_y > 0\}|$ and the label space $\mathcal{Y}$ is such that $\min(d^+, d^-) \geq k$, $||\boldsymbol{\mu}^+_{a,y}||_2 = 1, ||\boldsymbol{\mu}^-_{a,y}||_2 = 1$ for $a \in [m], y \in \mathcal{Y}$.*

Assumption 2. *(Bounded angle between $\boldsymbol{\mu}$ and $\mathbf{y}$) Let $\phi(\mathbf{u}, \mathbf{v})$ denote the angle between any two vectors $\mathbf{u}, \mathbf{v}$ in a Euclidean space. We assume that $\phi(\boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+) \in [0, \pi/2 - c)$, $\phi(\boldsymbol{\mu}^-_{a,y}, \mathbf{y}^-) \in [0, \pi/2 - c)$ $\forall a \in [m]$, and $y \in \mathcal{Y}_s$, for some sufficiently small $c \in (0, \pi/4]$ such that $\sin(c) \geq \max(\epsilon_0(d^+), \epsilon_0(d^-))$, where $\epsilon_0(d)$ is defined for some $n > n_0$ samples in* (6)*.*

Assumption 3. *σ is such that the recovery error with model* (4) *is at least as large as with* (3) *.*

These enable providing guarantees on recovering the mean vector magnitudes (1) and signs (2) and simplify the analysis (1), (3); all three can be relaxed at the expense of a more complex analysis.

Our first theoretical result shows that we have near-consistency in estimating the mean parameters in (3). We use standard notation $\tilde{\mathcal{O}}$ ignoring 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 $\mathbb{R}^{d^+,d^-}$). Suppose Assumptions 1 and 2 are met, a sufficiently large number of samples n are drawn from the model in* (3)*, and $k = |\mathcal{Y}_s|$. Then there exists a constant $C_0 > 0$ such that with high probability $\forall a \in [m]$ and $y \in \mathcal{Y}_s$,*

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

where

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

We interpret Theorem 1. 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.

4 Generalization Error for Structured Prediction in Finite Metric Spaces

We have access to labeling function outputs $\lambda_{a,i}, \ldots, \lambda_{m,i}$ for points x_i and noise rate estimates $\hat{\theta}_a, \ldots, \hat{\theta}_m$. How can we use these to infer unobserved labels y in (2)? Our approach is based on [NDRT13, vRW18],where the underlying loss function is modified to deal with noise. Analogously, we 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 $P(Y = y)$. We create pseudo-labels for each data point by drawing a random sample from the posterior distribution conditioned on the output of labeling functions: $\tilde{Z}^{(i)} = \tilde{z}_i \sim Y|\lambda_a = \lambda_a^{(i)}, \ldots, \lambda_m = \lambda_m^{(i)}; \hat{\boldsymbol{\theta}}$. We thus observe $(x_1, \tilde{z}_1), \ldots, (x_n, \tilde{z}_n)$ where $\tilde{z}_i$ is sampled as above. To overcome the effect of noise we create a perturbed version of the distance function using the noise rates, generalizing [NDRT13]. This requires us to characterize the noise distribution induced by our inference procedure. In particular we seek the probability that $\tilde{Z} = y_j$ when the true label is y_j. This can be expressed as follows. Let $\mathcal{Y}^m$ denote the m-fold Cartesian product of $\mathcal{Y}$ and let $\Lambda_u = (\lambda_1^{(u)}, \ldots, \lambda_m^{(u)})$ denote its u^{th} entry. We write

$$\mathbf{P}_{ij} = P_{\boldsymbol{\theta}}(\tilde{Z} = y_j|Y = y_i) = \sum_{u=1}^{|\mathcal{Y}^m|} P_{\boldsymbol{\theta}}(Y = y_j|\Lambda = \Lambda^{(u)}) \cdot P_{\boldsymbol{\theta}}(\Lambda = \Lambda^{(u)}|Y = y_i). \tag{7}$$

We define $\mathbf{Q}_{ij} = P_{\hat{\boldsymbol{\theta}}}(\tilde{Z} = y_j|Y = y_i)$ using $\hat{\boldsymbol{\theta}}$. $\mathbf{P}$ is the noise distribution induced by the true parameters $\boldsymbol{\theta}$ and $\mathbf{Q}$ is an approximation obtained from 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):

$$\tilde{d}_q(T, \tilde{Y} = y_j) := \sum_{i=1}^{k} (\mathbf{Q}^{-1})_{ji} d^2_{\mathcal{Y}}(T, Y = y_i) \quad \forall y_j \in \mathcal{Y}, \tag{8}$$

$$\tilde{F}_q(x, y) := \frac{1}{n} \sum_{i=1}^{n} \alpha_i(x) \tilde{d}_q(y, \tilde{z}_i) \qquad \hat{f}_q(x) = \arg\min_{y \in \mathcal{Y}} \tilde{F}_q(x, y). \tag{9}$$

We similarly define $\tilde{d}_p, \tilde{F}_p, \hat{f}_p$ using the true noise distribution $\mathbf{P}$. The perturbed distance $\tilde{d}_p$ is an unbiased estimator of the true distance. However we do not know the true noise distribution $\mathbf{P}$ hence we cannot use it for prediction. Instead we use $\tilde{d}_q$. Note that $\tilde{d}_q$ is no longer an unbiased estimator—its bias can be expressed as function of the parameter recovery error bound in Theorem 1.

4.2 Bounding the Generalization Error

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.

Our analysis uses 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 4. *(Noise model is not arbitrary) The true parameters $\boldsymbol{\theta}$ are such that $\sigma_{\min}(\mathbf{P}) > 0$, and the condition number $\kappa(\mathbf{P})$ is sufficiently small.*

Assumption 5. *(Normalized features) $|\alpha(x)| \leq 1$, for all $x \in \mathcal{X}$.*

Assumption 6. *(Proxy strong convexity) The function F in (2) satisfies the following property with some $\beta > 0$. 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:*

$$F\big(x, f(x)\big) \geq F\big(x, \hat{f}(x)\big) + \beta \cdot d^2_{\mathcal{Y}}\big(f(x), \hat{f}(x)\big) \qquad \forall x \in \mathcal{X}, \forall f \in \mathcal{F}. \tag{10}$$

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 (9)) be the minimizer over the noisy labels obtained from inference in Algorithm 1. Suppose Assumptions 4,5,6 hold. Then for $\epsilon_2 = k^{5/2} \cdot \tilde{\mathcal{O}}(\epsilon(d^+) + \epsilon(d^-)) \cdot \Big(1 + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big)$ and $c_1 = 1 + \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}$, with high probability,*

$$R(\hat{f}_q) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}) + \tilde{\mathcal{O}}\Big(\frac{c_1}{\beta} n^{-\frac{1}{2}}\Big) + \tilde{\mathcal{O}}\Big(\frac{3\epsilon_2}{\beta} n^{-\frac{1}{2}}\Big). \tag{11}$$

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 (perfect recovery) then we obtain the rate $\mathcal{O}(n^{-1/4})$, the same as in the case of access to clean 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 only pay an extra additive factor $\mathcal{O}(\sqrt{k}/d)$ in the excess risk when using pseudolabels.

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 $\exp : T_p\mathcal{M} \mapsto \mathcal{M}$ takes tangent vectors to manifold points. It enables switching between these tangent vectors: $\exp_p(v) = q$ implies that $d_{\mathcal{M}}(p, q) = \|v\|$. The logarithmic map operation $\log : \mathcal{M} \mapsto T_p\mathcal{M}$ takes manifold points to tangent vectors. Further, $\exp_p(v) = q$ is equivalent to $\log_p(q) = v$.

Invariant Our first contribution is a simple invariant that enables us to recover the error parameters. Note that we cannot rely on the finite metric-space technique, since the manifolds we consider have an infinite number of points. 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 preservation of concentration when moving from distribution $\cosh(d)$ to d^2. Then,

Theorem 3. *Let $\mathcal{M}$ be a hyperbolic manifold. Fix $0 < \delta < 1$ and let $\Delta(\delta) = \min_\rho Pr\Big(\forall i, d_{\mathcal{Y}}(\lambda_{a,i}, \lambda_{b,i}) \leq \rho\Big) \geq 1 - \delta$. Then, there exists a constant C_1 so that with probability at least $1 - \delta$, $\mathbb{E}|\hat{\mathbb{E}} d^2_{\mathcal{Y}}(\lambda_a, y)) - \mathbb{E} d^2_{\mathcal{Y}}(\lambda_a, y)| \leq C_1 \cosh(\Delta(\delta))^{3/2} / C_0 \sqrt{2n}$.*

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). In addition, it is possible to replace the $\Delta(\delta)$ term by applying a version of McDiarmid's inequality for unbounded spaces as in [Kon14].

Next, we adapt the downstream model predictor (2) in the following way. Let $\hat{\mu}_a^2 = \hat{\mathbb{E}}[d^2_{\mathcal{Y}}(\lambda_a, y)]$. 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

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

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

Assumption 7. *(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 $k_q^b(a) = 1 - (\|\log_q(a) - \log_q(b)\|^2 - d^2_{\mathcal{Y}}(a, b))/d^2_{\mathcal{Y}}(q, b)$. We assume that $k_q^b(a)$ is lower bounded by $k_{\min}$.*

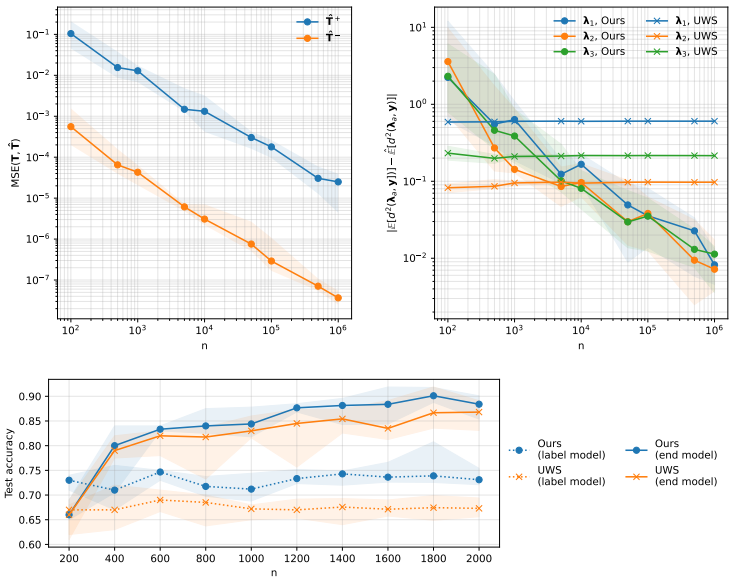

Figure 2: **Finite metric space case.** Parameter estimation improves with samples n in learning to rank—showing nearly-consistent behavior. Our tensor decomposition estimator outperforms [SLV$^+$22]. In particular, (top left) as the number of samples increases, our estimates of the positive and negative components of $\mathbf{T}$ improve. (Top right) the improvements in $\mathbf{T}$ recovery with more samples translates to significantly improved performance over [SLV$^+$22], which is close to constant across n. (Bottom) this improved parameter estimation further translates to improvements in label model accuracy (using only the noisy estimates for prediction, without training an end model) and end model generalization. For the top two plots, we use $\boldsymbol{\theta} = [6, 3, 8]$, and in the bottom plot, we use $\boldsymbol{\theta} = [0, 0, 1]$. In all plots, we report medians along with upper and lower quartiles across 10 trials.

Assumption 8. *(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 such that,*

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

Note that as n grows, 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.

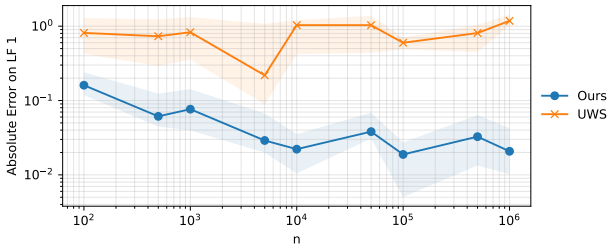

Figure 3: **Continuous case.** Parameter estimation improves with more samples in the hyperbolic regression problem. Our estimator outperforms [SLV$^+$22]. Here, we use different randomly sampled values of $\boldsymbol{\theta}$ for each run. We report medians along with upper and lower quartiles across 10 trials.

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 so for symmetric Riemannian manifolds using the tools in [AH91].

6 Experiments

Finally, we validate our theoretical claims with experimental results demonstrating improved parameter recovery and end model generalization using our techniques over that of prior work [SLV$^+$22]. We illustrate both the finite metric space and continuous space cases by targeting rankings (i.e., permutations) and hyperbolic spaces. In the case of rankings we show that our pseudo-Euclidean embeddings with tensor decomposition estimator yields stronger parameter recovery and downstream generalization than [SLV$^+$22]. In the case of hyperbolic regression (an example of a Riemannian manifold), we show that our estimator yields improved parameter recovery over [SLV$^+$22].

Finite metric spaces: Learning to rank To experimentally evaluate our tensor decomposition estimator for finite metric spaces, we consider the problem of learning to rank. We construct a synthetic dataset whose ground truth comprises n samples of two distinct rankings among the finite metric space of all length-four permutations. We construct three labeling functions by sampling rankings according to a Mallows model, for which we obtain pseudo-Euclidean embeddings to use with our tensor decomposition estimator.

In Figure 2 (top left), we show that as we increase the number of samples, we can obtain an increasingly accurate estimate of $\mathbf{T}$—exhibiting the *nearly-consistent* behavior predicted by our theoretical claims. This leads to downstream improvements in parameter estimates, which also become more accurate as n increases. In contrast, we find that the estimates of the same parameters given by [SLV$^+$22] do not improve substantially as n increases, and are ultimately worse (see Figure 2, top right). Finally, this leads to improvements in the label model accuracy as compared to that of [SLV$^+$22], and translates to improved accuracy of an end model trained using synthetic samples (see Figure 2, bottom).

Riemannian manifolds: Hyperbolic regression We similarly evaluate our estimator using synthetic labels from a hyperbolic manifold, matching the setting of Section 5. As shown in Figure 3, we find that our estimator consistently outperforms that of [SLV$^+$22], often by an order of magnitude.

7 Conclusion

We studied the theoretical properties of weak supervision applied to structured prediction in two general scenarios: label spaces that are finite metric spaces or constant-curvature manifolds. We introduced ways to estimate the noise rates of labeling functions, achieving consistency or near-consistency. Using these tools, we established that with suitable modifications downstream structured prediction models maintain generalization guarantees. Future directions include extending these results to even more general manifolds and removing some of the assumptions needed in our analysis.

Acknowledgments

We are grateful for the support of the NSF (CCF2106707), the American Family Funding Initiative and the Wisconsin Alumni Research Foundation (WARF). We are thankful to Changho Shin and Harshavardhan Adepu for the discussions and feedback.

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.

[BDDH14] Richard Baraniuk, Mark A Davenport, Marco F Duarte, and Chinmay Hegde. An introduction to compressive sensing. 2014.

[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.

[Kon14] Aryeh Kontorovich. Concentration in unbounded metric spaces and algorithmic stability. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 28–36, Bejing, China, 22–24 Jun 2014. PMLR.

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

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

[LRBM06] Julian Laub, Volker Roth, Joachim M Buhmann, and Klaus-Robert Müller. On the information and representation of non-euclidean pairwise data. *Pattern Recognition*, 39(10):1815–1826, 2006.

[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.

[PHD+06] Elżbieta Pękalska, Artsiom Harol, Robert P. W. Duin, Barbara Spillmann, and Horst Bunke. Non-euclidean or non-metric measures can be informative. In Dit-Yan Yeung, James T. Kwok, Ana Fred, Fabio Roli, and Dick de Ridder, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, pages 871–880, 2006.

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

[PPD01] Elżbieta Pękalska, Pavel Paclik, and Robert P.W. Duin. A generalized kernel approach to dissimilarity-based classification. *Journal of Machine Learning Research*, 2:175–211, 2001.

[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
$\mathcal{Y}_s$	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
$\theta_a, \hat{\theta}_a$	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	proxy 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 (9) 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 $\mathbf{P}$
$\phi(\mathbf{u}, \mathbf{v})$	angle between vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$

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

B Proofs for Parameter Estimation Error in Discrete Spaces

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 mutually independent labeling functions triplet, as defined in* (5) *using a sufficiently large number n of i.i.d observations drawn from models in equation* (4)*. Suppose there are sufficiently many such triplets to cover all labeling functions. Let $\hat{\boldsymbol{\mu}}^+_{a,y}, \hat{\boldsymbol{\mu}}^-_{a,y}$ be the estimated parameters returned by the algorithm 1 for all $a \in [m]$. Let $\epsilon(d)$ be defined as above in equation* (6)*, then the following holds with high probability for all labeling functions,*

$$||\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}||_2 \leq \mathcal{O}(\epsilon(d^+)) \quad \textit{and} \quad ||\boldsymbol{\mu}^-_{a,y} - \hat{\boldsymbol{\mu}}^-_{a,y}||_2 \leq \mathcal{O}(\epsilon(d^-)) \quad \forall a \in [m]\ \forall y \in \mathcal{Y}_s \tag{12}$$

Proof. The result follows by first showing that our setting and assumptions imply that the conditions of Theorems 1 and 5 in [AGJ14] are satisfied, which allows us to adopt their results. We then translate the result in order to state it in terms of the ℓ_2 distance.

The tensor concentration result in Theorem 1 in [AGJ14] relies heavily on the noise matrices satisfying the Restricted Isometry Property (RIP) property. The authors make an explicit assumption that the noise model satisfies this condition. In our setting, we have a specific form of the noise model that allows us to show that this assumption is satisfied. The RIP condition is satisfied for sub-Gaussian noise matrices [BDDH14]. Our noise matrices are supported on a discrete space and have bounded entries, and so are sub-Gaussian.

The other required conditions on the norms of factor matrices and the number of latent factors are implied by Assumption 1. Thus, we can adopt the results on recovery of parameters $\boldsymbol{\mu}_{a,y}$ and the prior weights w_y from [AGJ14]. The result gives us for all $a \in [m], y \in \mathcal{Y}_s$,

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

and

$$|w_y - \hat{w}_y| \leq \mathcal{O}\Big(\max\big(\epsilon(d^+), \epsilon(d^-)\big)/k\Big),$$

where $\text{dist}(\mathbf{u}, \mathbf{v})$ is defined as follows. For any $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$,

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

Next, we translate the result to the Euclidean distance. For $\mathbf{u}, \mathbf{v} \in \mathbb{R}^d$ with $||\mathbf{u}||, ||\mathbf{v}|| = 1$, it is easy to see that

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

This notion of distance is oblivious to sign recovery. However, when sign recovery is possible then the Euclidean distance can be bounded as follows,

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

Next we make use of Assumption 2 to recover the signs of $\boldsymbol{\mu}^+, \boldsymbol{\mu}^-$. The assumption bounds the angle between true $\boldsymbol{\mu}^+_{a,y}$ and $\mathbf{y}^+$ between $[0, \pi/2 - c)$ for some sufficiently small $c \in (0, \pi/4]$ such that $\sin(c) > \max(\epsilon_0(d^+), \epsilon_0(d^-))$, where $\epsilon_0(d)$ is defined for some $n_0 < n$ samples in equation (6). We measure $\phi(\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+)$ and $\phi(-\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+)$ and claim that whichever makes an acute angle with $\mathbf{y}^+$ has the correct sign.

We have that $\phi(\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+) \leq \phi(\hat{\boldsymbol{\mu}}^+_{a,y}, \boldsymbol{\mu}^+_{a,y}) + \phi(\boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+)$. Let $s \in \{-1, +1\}$ be the correct sign, then,

$$\begin{aligned}
\phi(s\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+) &\leq \phi(s\hat{\boldsymbol{\mu}}^+_{a,y}, \boldsymbol{\mu}^+_{a,y}) + \phi(s\boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+) \\
&\leq \sin^{-1}(\epsilon(d^+)) + \pi/2 - c \\
&< \pi/2 - (\sin^{-1}(\max(\epsilon_0(d^+), \epsilon_0(d^-))) - \sin^{-1}(\epsilon(d^+))) \\
&< \pi/2 \quad \text{since } \sin^{-1} \text{ is an increasing function in the domain under consideration.}
\end{aligned}$$

With the correct sign $\sin^{-1}(\epsilon(d^+)) < \pi/2$ and so is $\phi(s\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+)$. Thus with incorrect sign $\phi(-s\hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+) > \pi/2$.

Hence, after disambiguating the signs we have,

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

and similarly for $\boldsymbol{\mu}^-_{a,y}$. Next with n, d sufficiently large such that $\epsilon(d^+), \epsilon(d^-) \leq 1$, 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 $\mathbb{R}^{d^+,d^-}$). Suppose Assumptions 1 and 2 are met, a sufficiently large number of samples n are drawn from the model in* (3)*, and $k = |\mathcal{Y}_s|$. Then there exists a constant $C_0 > 0$ such that with high probability $\forall a \in [m]$ and $y \in \mathcal{Y}_s$,*

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

where

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

Proof. We prove this by using the bounds on errors in the estimates of $\boldsymbol{\mu}^+_{a,y}$ and $\boldsymbol{\mu}^-_{a,y}$ from Lemma 2. We proceed by bounding the errors in two parts for $\mathbb{E}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y}^+)]$ and $\mathbb{E}[d^2_\phi(\boldsymbol{\lambda}^-_a, \mathbf{y}^-)]$ separately and then combine them to get the bound on overall parameter estimation error.

We first bound the error for $\mathbb{E}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y}^+)]$. The true mean parameter (i.e., the true expected squared distance) can be expanded as follows:

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

The estimate $\hat{\mathbb{E}}_{\boldsymbol{\lambda}}[||\boldsymbol{\lambda}^+_a||^2_2]$ is computed empirically. The first term is estimated observed LF outputs, i.e. $\hat{\mathbb{E}}_{\boldsymbol{\lambda}}[||\boldsymbol{\lambda}^+_a||^2_2] = \frac{1}{n}\sum_{i=1}^n ||\boldsymbol{\lambda}^{(i),+}_a||^2_2$. The second term is computed by using the estimated prior on the labels and for the last term we plug in the estimate of $\boldsymbol{\mu}^+_{a,y}$ computed using the tensor-decomposition algorithm. Putting them all together we have the following estimator:

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

We want to bound the error of our estimator i.e. the difference $|\mathbb{E}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y})] - \hat{\mathbb{E}}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y})]|$. For this first consider the following,

$$\begin{aligned} |\mathbb{E}_{\mathbf{y}}\langle \boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+ \rangle - \hat{\mathbb{E}}_{\mathbf{y}}\langle \hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+ \rangle| &= \sum_y \Big| \big\langle \big(w_y \boldsymbol{\mu}^+_{a,y} - \hat{w}_y \hat{\boldsymbol{\mu}}^+_{a,y}\big), \mathbf{y}^+ \big\rangle \Big| \\ &\leq \sum_y |w_y \langle \big(\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}\big), \mathbf{y} \rangle| + \sum_y \mathcal{O}(\epsilon(d^+)/k) |\langle \hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y} \rangle| \\ &\leq \sum_y |w_y \langle \big(\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}\big), \mathbf{y} \rangle| + \mathcal{O}(\epsilon(d^+)) \\ &\leq \sum_y w_y ||\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}||_2 ||\mathbf{y}||_2 + \mathcal{O}(\epsilon(d^+)) \\ &\leq \mathcal{O}(\epsilon(d^+)). \end{aligned}$$

Here we used $||\boldsymbol{\mu}^+_{a,y} - \hat{\boldsymbol{\mu}}^+_{a,y}||_2 \leq \mathcal{O}(\epsilon(d^+))$ and $||\boldsymbol{\mu}^+_{a,y}||_2, ||\hat{\boldsymbol{\mu}}^+_{a,y}||_2 = 1, ||\mathbf{y}^+||_2 \leq 1, ||\boldsymbol{\lambda}^+_a||^2_2 \leq 1$ and $|w_y - \hat{w}_y| \leq \mathcal{O}(d^+)/k$. Hence the parameter estimator error,

$$\begin{aligned} \Big| \mathbb{E}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y})] - \hat{\mathbb{E}}[d^2_\phi(\boldsymbol{\lambda}^+_a, \mathbf{y})] \Big| &\leq \Big| \mathbb{E}_{\boldsymbol{\lambda}}[||\boldsymbol{\lambda}^+_a||^2_2 - \hat{\mathbb{E}}_{\boldsymbol{\lambda}}[||\boldsymbol{\lambda}^+_a||^2_2] \Big| + 2|\mathbb{E}_{\mathbf{y}}\langle \boldsymbol{\mu}^+_{a,y}, \mathbf{y}^+ \rangle - \hat{\mathbb{E}}_{\mathbf{y}}\langle \hat{\boldsymbol{\mu}}^+_{a,y}, \mathbf{y}^+ \rangle| \\ &\leq \mathcal{O}(1/\sqrt{n}) + \mathcal{O}(\epsilon(d^+)) \\ &\leq \mathcal{O}(\epsilon(d^+)). \end{aligned}$$

In the second step, we bound the first term by $\mathcal{O}(1/\sqrt{n})$ via standard concentration inequalities. Doing the same calculations for $\boldsymbol{\lambda}_a^-$, we obtain

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

The overall error in mean parameters is then

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

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:

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

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 $C_0 = \max_{a\in[m]} e_{\min}(A_a)$ concludes the proof. □

C Proofs for Generalization Error in Discrete Space

In this section we give the proof for the generalization error bound in the discrete label spaces. We first show that the perturbed (noise-aware) distance function $\tilde{d}_p$ is an unbiased estimator of the true distance. Using this we show that the noise aware score function $\tilde{F}_p$ is a good uniform approximation of the score function F. Then we show that the minimizer $\hat{f}_p$ of $\tilde{F}_p$ is close to the minimizer $\hat{f}$ and that this closeness depends on how well $\tilde{F}_p$ approximates F. Next, showing that $\tilde{F}_q$ is a good uniform approximation of $\tilde{F}_p$ using the results from previous section on parameter recovery leads to the result on generalization error of $\hat{f}_q$.

Lemma 3. *Let the distribution $\tilde{Y}|Y$ be given by $\mathbf{P}$ a $k \times k$ transition probability matrix with $\mathbf{P}_{ij} = \mathbb{P}(\tilde{Y} = y_j | Y = y_i)$ and suppose $\mathbf{P}$ is invertible. Let the pseudo-distance $\tilde{d}_p$ be defined as in* (8) *then,*

$$\mathbb{E}_{\tilde{Y}|Y=y_i}\big[\tilde{d}_p(T, \tilde{Y})\big] = d_{\mathcal{Y}}^2(T, y_i). \tag{13}$$

Proof. Set $\tilde{\mathbf{d}}_p \in \mathbb{R}^k$ with ith entry $\tilde{\mathbf{d}}_p[i]$ given by $\tilde{d}_p(T, \tilde{Y} = y_i)$ and similarly define $\mathbf{d}$ with $\mathbf{d}[i] = d_{\mathcal{Y}}^2(T, y_i)$. Then we note that $\tilde{\mathbf{d}}_p$ satisfies the following,

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

and we are done. □

Next, we show that the noisy score function $\tilde{F}_p$ concentrates around the true score function F for all x and y with high probability.

Lemma 4. *Let F and $\tilde{F}_p$ be defined as in* (9) *and* (2) *over n i.i.d. samples. Then the following holds for any $x \in \mathcal{X}, y \in \mathcal{Y}$ with high probability,*

$$|F(x, y) - \tilde{F}_p(x, y)| \leq \tilde{\mathcal{O}}\Big(\big(1 + \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}\big)\sqrt{\frac{1}{n}}\Big) \quad \forall x \in \mathcal{X}, \forall y \in \mathcal{Y}_s, \tag{14}$$

where $\sigma_{\min}(\mathbf{P})$ is the minimum singular value of $\mathbf{P}$.

Proof. 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 the true noise model $\mathbf{P}$ be $\tilde{y}_i$. Recall the definitions of the score functions F and $\tilde{F}_p$ for any $x \in \mathcal{X}$ and y in $\mathcal{Y}$,

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

Taking their difference,

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

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 $\mathbf{P}$. From Lemma 3 we have $\mathbb{E}_{\tilde{Y}|Y=y_i}[\xi(y, y_i, \tilde{Y})] = 0$ and 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 random variables and $\alpha_i(x) \leq 1$. The ξ are bounded as follows as long as the spectral decomposition of $\mathbf{P}$ is not arbitrary,

$$\max_{z \in \mathcal{Y}_s} \tilde{d}_p(y,z) = ||\tilde{\mathbf{d}}_p||_\infty = ||\mathbf{P}^{-1}\mathbf{d}||_\infty \leq ||\mathbf{P}^{-1}||_\infty ||\mathbf{d}||_\infty.$$

Now using the fact that $||\mathbf{d}||_\infty \leq 1$ and properties of matrix norms we get,

$$||\mathbf{P}^{-1}||_\infty ||\mathbf{d}||_\infty \leq ||\mathbf{P}^{-1}||_\infty \leq \sqrt{k}||\mathbf{P}^{-1}||_2 \leq \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}.$$

Moreover, $\forall y, z \in \mathcal{Y}_s, d^2_{\mathcal{Y}}(y,z) \leq 1$ which gives us the magnitude of random variables $\xi(y,z,\tilde{z})$ is upper bounded by $c_1 := 1 + \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}$ $\forall y, z, \tilde{z} \in \mathcal{Y}_s$. Thus using Hoeffding's inequality and union bound over all $y \in \mathcal{Y}_s$ we get,

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

Note that, the statement holds for $x \in \mathcal{X}$ without requiring an explicit union bound over x. It is because the above concentration depends only on the labels and the events that the above inequality does not hold for any distinct $x_1, x_2 \in \mathcal{X}$ are the same. □

Now, we show that the distance between minimizer of $\tilde{F}_p$ and F is bounded.

Lemma 5. *Let $\hat{f}$ be the minimizer as defined in* (2) *over the clean labels and let $\hat{f}_p$ (defined in eq.* (9)*) 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* (1)*, then with high probability,*

$$d^2_{\mathcal{Y}}\big(\hat{f}_p(x), \hat{f}(x)\big) \leq \tilde{\mathcal{O}}\Big(\frac{c_1}{\beta}\sqrt{\frac{1}{n}}\Big) \quad \forall x \in \mathcal{X}. \tag{15}$$

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 $t = \tilde{\mathcal{O}}\Big(c_1\sqrt{\frac{1}{n}}\Big)$,

$$F\big(x, f(x)\big) - t \leq \tilde{F}_p\big(x, f(x)\big) \leq F\big(x, f(x)\big) + t \quad \forall f : \mathcal{X} \mapsto \mathcal{Y}_s.$$

From Assumption 6 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 sufficiently 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 $\tilde{F}_p$,

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

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'. □

Next we show that a good estimate of true noise matrix $\mathbf{P}$ by $\mathbf{Q}$ leads to $\tilde{F}_q$ being uniformly close to $\tilde{F}_p$.

Lemma 6. *Let* $\mathbf{Q}$, $\mathbf{P}$ *be the distributions defined in equation* (7)*, and* $\tilde{d}_q(T, \tilde{Y})$ *be the distance function as in* (8)*, if* $\max_{ij} |\mathbf{P}_{ij} - \mathbf{Q}_{ij}| = \epsilon$*,*

$$\big|\tilde{d}_q(y, \tilde{z}_i) - \tilde{d}_p(y, \tilde{z}_i)\big| \leq \mathcal{O}\Big(k^2\Big(\sigma_{\max}(\mathbf{P}) + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big) \cdot \epsilon\Big) \qquad \forall y \in \mathcal{Y}_s. \tag{16}$$

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 $\tilde{\mathbf{d}}_p \in \mathbb{R}^k$ with $\tilde{\mathbf{d}}_p[i] = \tilde{d}_p(T, \tilde{Y} = y_i)$, 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 $\mathbf{P}$,

$$\tilde{\mathbf{d}}_q - \tilde{\mathbf{d}}_p = \mathbf{Q}^{-1}\mathbf{d} - \mathbf{P}^{-1}\mathbf{d} = \big(\mathbf{Q}^{-1} - \mathbf{P}^{-1}\big)\mathbf{d}.$$

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

$$\begin{aligned}
||\tilde{\mathbf{d}}_p - \tilde{\mathbf{d}}_q||_\infty &\leq ||(\mathbf{P} + \Delta\mathbf{P})^{-1} - \mathbf{P}^{-1}||_\infty ||\mathbf{d}||_\infty, \\
&\leq \sqrt{k}||(\mathbf{P} + \Delta\mathbf{P})^{-1} - \mathbf{P}^{-1}||_2 ||\mathbf{d}||_\infty, \\
&= \sqrt{k}\Big(\kappa(\mathbf{P})||\mathbf{P}^{-1}||_2||\Delta\mathbf{P}||_2\Big) + \sqrt{k}\mathcal{O}(||\Delta\mathbf{P}||_2^2), \\
&\leq \sqrt{k} \cdot \kappa(\mathbf{P})||\mathbf{P}^{-1}||_2 \cdot \epsilon k + \mathcal{O}(\epsilon^2 k^{5/2}), \\
&\leq \mathcal{O}\Big(k^{5/2}\Big(1 + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big) \cdot \epsilon\Big) =: c_2.
\end{aligned}$$

□

Lemma 7. *For $\tilde{F}_p$ and $\tilde{F}_q$ defined in* (9) *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.*

$$|\tilde{F}_p(x, y) - \tilde{F}_q(x, y)| \leq \tilde{\mathcal{O}}\Big((2c_1 + c_2)\sqrt{\frac{1}{n}}\Big) \qquad \forall y \in \mathcal{Y}_s, \forall x \in \mathcal{X}. \tag{17}$$

with $c_2 = k^{5/2} \cdot \epsilon \cdot \Big(1 + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big)$ *and* $c_1 = 1 + \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}$,

Proof. Recall, random variables $\tilde{Y}$,$\tilde{Z}$ denote the noisy labels drawn from true and estimated noise distributions $\mathbf{P}, \mathbf{Q}$ respectively and $\tilde{y}_i, \tilde{z}_i$ denote their draw for data point x_i. Note that we do not know $\mathbf{P}$ and $\tilde{y}_i$ in practice and we only know $\mathbf{Q}, \tilde{z}_i$. Here we are using $\mathbf{P}$ and $\tilde{y}_i$ to compare our actual estimates using samples $\tilde{z}_i$ against the estimates one could have obtained from $\tilde{y}_i$.

Recall the definitions,

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

Then,

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

Thus,

$$\begin{aligned} \mathbb{E}_{\tilde{Y}, \tilde{Z}|Y=y_i}\big[\tilde{d}_p(y, \tilde{Y}) - \tilde{d}_q(y, \tilde{Z})\big] &= \mathbb{E}_{\tilde{Z}|Y=y_i}[\tilde{d}_q(y, \tilde{Y})] - \mathbb{E}_{\tilde{Z}|Y=y_i}[\tilde{d}_q(y, \tilde{Z})] \\ &= d^2_{\mathcal{Y}}(y, y_i) - d^2_{\mathcal{Y}}(y, y_i) = 0 \end{aligned}$$

Finally $\mathbb{E}_{\tilde{Y}, \tilde{Z}}[\xi(y, \tilde{Y}, \tilde{Z})] = 0$.

Next,

$$\begin{aligned} \tilde{d}_p(y, \tilde{y}_i) - \tilde{d}_q(y, \tilde{z}_i) &\leq |\tilde{d}_p(y, \tilde{y}_i) - \tilde{d}_q(y, \tilde{z}_i)| \\ &\leq |\tilde{d}_p(y, \tilde{y}_i) - \tilde{d}_p(y, \tilde{z}_i) + \tilde{d}_p(y, \tilde{z}_i) - \tilde{d}_q(y, \tilde{z}_i)| \\ &\leq |\tilde{d}_p(y, \tilde{y}_i) - d^2_{\mathcal{Y}}(y, \tilde{z}_i) + d^2_{\mathcal{Y}}(y, \tilde{z}_i) - \tilde{d}_p(y, \tilde{z}_i) + \tilde{d}_p(y, \tilde{z}_i) - \tilde{d}_q(y, \tilde{z}_i)| \\ &\leq |\tilde{d}_p(y, \tilde{y}_i) - d^2_{\mathcal{Y}}(y, \tilde{z}_i)| + |d^2_{\mathcal{Y}}(y, \tilde{z}_i) - \tilde{d}_p(y, \tilde{z}_i)| + |\tilde{d}_p(y, \tilde{z}_i) - \tilde{d}_q(y, \tilde{z}_i)| \\ &\leq 2c_1 + |\tilde{d}_p(y, \tilde{z}_i) - \tilde{d}_q(y, \tilde{z}_i)| \\ &\leq 2c_1 + c_2. \end{aligned}$$

The first two terms are upper bounded as in Lemma 4 and the last term is bounded using Lemma 6. Since $\alpha_i(x) \leq 1$ and $|\xi(y, \tilde{y}_i, \tilde{z}_i)|$ are upper bounded by $2c_1 + c_2$ as shown above, we have that $|\alpha_i(x) \cdot \xi(y, \tilde{y}_i, \tilde{z}_i)| \leq 2c_1 + c_2$. □

Lemma 8. *Let $\hat{f}_p$ be the minimizer as defined in* (9) *over the noisy labels drawn from* $\mathbf{P}$*, and let $\hat{f}_q$ (defined in eq.* (9)*) be the minimizer over the noisy labels obtained from conditional distribution* $\mathbf{Q}$*. Then with high probability,*

$$d^2_{\mathcal{Y}}\big(\hat{f}_q(x), \hat{f}(x)\big) \leq \tilde{\mathcal{O}}\Big(\frac{1}{\beta}(3c_1 + c_2)\sqrt{\frac{1}{n}}\Big) \qquad \forall x \in \mathcal{X}. \tag{18}$$

Proof. Let $t_1 = \tilde{\mathcal{O}}\Big(c_1\sqrt{\frac{1}{n}}\Big)$ and $t_2 = \tilde{\mathcal{O}}\Big((2c_1 + c_2)\sqrt{\frac{1}{n}}\Big)$, 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. □

The following lemmas bound the estimation error between noise matrices $\mathbf{P}$ and $\mathbf{Q}$ using the estimation error in the canonical parameters.

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 $\mathbf{d}_a^{(u,i)} = d_{\mathcal{Y}}^2(\Lambda_a^u, y_i)$

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

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,

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

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

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

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 (7) 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.

Finally, we restate and prove our generalization error 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* (9)*) be the minimizer over the noisy labels obtained from inference in Algorithm 1. Suppose Assumptions 4,5,6 hold. Then for $\epsilon_2 = k^{5/2} \cdot \tilde{\mathcal{O}}(\epsilon(d^+) + \epsilon(d^-)) \cdot \Big(1 + \frac{\kappa(\mathbf{P})}{\sigma_{\min}(\mathbf{P})}\Big)$ and $c_1 = 1 + \frac{\sqrt{k}}{\sigma_{\min}(\mathbf{P})}$, with high probability,*

$$R(\hat{f}_q) \leq R(f^*) + \mathcal{O}(n^{-\frac{1}{4}}) + \tilde{\mathcal{O}}\Big(\frac{c_1}{\beta} n^{-\frac{1}{2}}\Big) + \tilde{\mathcal{O}}\Big(\frac{3\epsilon_2}{\beta} n^{-\frac{1}{2}}\Big). \tag{11}$$

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.

□

D 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, λ_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 and their distributions are symmetric. 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 $\Delta(\delta) = \min_\rho Pr\Big(\forall i, d_{\mathcal{Y}}(\lambda_{a,i}, \lambda_{b,i}) \leq \rho\Big) \geq 1 - \delta$. Then, there exists a constant C_1 so that with probability at least $1 - \delta$, $\mathbb{E}|\hat{\mathbb{E}}d^2_{\mathcal{Y}}(\lambda_a, y)) - \mathbb{E}d^2_{\mathcal{Y}}(\lambda_a, y)| \leq C_1 \cosh(\Delta(\delta))^{3/2}/C_0\sqrt{2n}$.*

Proof. [Kon14] First, we will condition on the event that the observed outputs have maximal distance (i.e., diameter) $\Delta(\delta)$. 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

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

Integrating the expression above in t, we obtain

$$\mathbb{E}|\hat{\mathbb{E}}\cosh(d(\lambda_a, \lambda_b)) - \mathbb{E}\cosh(d(\lambda_a, \lambda_b))| \leq \frac{\sqrt{\pi\cosh(\Delta(\delta))}}{\sqrt{2n}}. \tag{19}$$

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

$$\hat{\mathbb{E}}\cosh(d(\lambda_a, y)) = \sqrt{\frac{\hat{\mathbb{E}}\cosh d(\lambda_a, \lambda_b)\hat{\mathbb{E}}\cosh d(\lambda_a, \lambda_c)}{(\hat{\mathbb{E}}\cosh d(\lambda_b, \lambda_c))^2}}.$$

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 $\hat{\nu}(a, b) \leq \cosh(\Delta(\delta))$. With this, we can begin our perturbation analysis. Applying Lemma 1, we have that

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

To see why the last step holds, note that $\sqrt{\hat{\nu}(a,c)} \leq \sqrt{\cosh(\Delta(\delta))}$, 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 $\mathbb{E}|\sqrt{\hat{\nu}(a,b)} - \sqrt{\nu(a,b)}| \leq \mathbb{E}|\hat{\nu}(a,b) - \nu(a,b)| \leq \frac{\sqrt{\pi \cosh(\Delta(\delta))}}{\sqrt{2n}}$ using (19).

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

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

The first expectation in the r.h.s is bounded using the same steps as above. The second expectation is bounded as follows,

$$\mathbb{E}\left|\sqrt{\frac{\nu(a,b)\nu(a,c)}{\hat{\nu}(b,c)^2}} - \sqrt{\frac{\nu(a,b)\nu(a,c)}{\nu(b,c)^2}}\right| \leq \mathbb{E}\left|\sqrt{\nu(a,b)\nu(a,c)}\Big(\frac{\big(\hat{\nu}(b,c) - \nu(b,c)\big)\big(\hat{\nu}(b,c) + \nu(b,c)\big)}{\hat{\nu}(b,c)\nu(b,c)}\Big)\right|$$

Here, the denominator is lower bounded by 1 and in the numerator $\sqrt{\nu(a,b)\nu(a,c)} \leq \cosh(\Delta(\delta))$ and $\hat{\nu}(b,c)+\nu(b,c) \leq 2\cosh(\Delta(\delta))$ and $\mathbb{E}(\hat{\nu}(b,c)-\nu(b,c)) \leq \frac{\sqrt{\pi\cosh(\Delta(\delta))}}{\sqrt{2n}}$. Putting it all together, with probability at least $1-\delta$,

$$\mathbb{E}|\hat{\mathbb{E}}\cosh(d(\lambda_a, y)) - \mathbb{E}\cosh(d(\lambda_a, y))| \leq \frac{2\sqrt{\pi}\cosh(\Delta(\delta)) + 2\sqrt{\pi}(\cosh(\Delta(\delta))^{3/2}}{\sqrt{n}}. \tag{20}$$

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,

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

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 7 and 8 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 the minimizer of the objective function under P' is not affected by uniformly scaling the distances by some constant. If we do so repeatedly, we can shrink the region in which this minimizer—and that of the objective function for P—are found. This means that the distance between the two minimizers must be arbitrarily small, so that by a limit argument, they must be the same.

□

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 such that,*

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

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.

Using the hugging function assumption, we have that,

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

We also have that

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

Then,

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

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

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

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

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

We apply Cauchy-Schwarz, obtaining

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

Since $\|\log_q(\hat{f}(x)\| = d_{\mathcal{Y}}(q, \hat{f}(x))$, we then have that

$$\alpha_{\min}(1-k_{\min}) d_{\mathcal{Y}}(q, \hat{f}(x)) \leq 2\alpha_{\max}\|\bar{q}\|.$$

Squaring both sides, we obtain

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

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

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

Thus we obtain

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

or

$$\mathbb{E} d_{\mathcal{Y}}^2(q, \hat{f}(x)) \leq 4\frac{\alpha_{\max}^2}{\alpha_{\min}^2} \frac{\sigma_o^2}{n(1-k_{\min})^2}. \tag{21}$$

We use the same approach, but apply it to the objective function that involves the n samples of the m LFs drawn from the distribution P'. In this case, the $\bar{q}$ vector becomes $\frac{1}{n}\sum_{i=1}^n(\sum_{a=1}^m \beta_a \log_q(\lambda_{a,i}))$.

Doing so yields

$$\alpha_{\min}^2(1-k_{\min})^2\mathbb{E}d_{\mathcal{Y}}^2(q,\tilde{f}(x)) \leq 4\alpha_{\max}^2\frac{\sum_{a=1}^m \beta_a^2\sigma_a^2}{n},$$

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

$$\alpha_{\min}^2(1-k_{\min})^2\mathbb{E}d_{\mathcal{Y}}^2(q,\tilde{f}(x)) \leq 8\alpha_{\max}^2\frac{\sum_{a=1}^m \beta_a^2(\sigma_o+\hat{\mu}_a^2)}{n},$$

or,

$$\mathbb{E}d_{\mathcal{Y}}^2(q,\tilde{f}(x)) \leq 8\frac{\alpha_{\max}^2}{\alpha_{\min}^2}\frac{\sum_{a=1}^m \beta_a^2(\sigma_o^2+\hat{\mu}_a^2)}{n(1-k_{\min})^2}. \tag{22}$$

Now, again using triangle inequality,

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

Plugging (22) and (21) into this bound produces the result. □

E Additional Details on Continuous Label Space

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 function Note that it is possible to derive a lower bound on the hugging function 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 function 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 $\mathrm{Var}[\frac{1}{m}\sum_{a=1}^m \beta_a\lambda_a]$ 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.

F Extended Background on Pseudo-Euclidean Embeddings

We provide some additional background on pseudo-metric spaces and pseudo-Euclidean embeddings. Our roadmap is as follows. First, we note that pseudo-Euclidean spaces are a particular kind of pseudo-metric space, so we provide additional background and formal definitions for these pseudo-metric spaces. Afterwards, we explain some of the ideas behind pseudo-Euclidean spaces, comparing them to standard Euclidean spaces in the context of embeddings.

F.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,*

$$\forall \mathbf{y}, \mathbf{z} \in \mathcal{Y} \qquad d_{\mathcal{Y}}(\mathbf{y}, \mathbf{z}) = d_{\mathcal{Y}}(\mathbf{y}, \mathbf{z}) \tag{23}$$

(Symmetry)

$$\forall \mathbf{y} \in \mathcal{Y} \qquad d_{\mathcal{Y}}(\mathbf{y}, \mathbf{y}) = 0 \tag{24}$$

(Reflexivity)

$$\forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{Y} \qquad d_{\mathcal{Y}}(\mathbf{y}, \mathbf{x}) \leq d_{\mathcal{Y}}(\mathbf{y}, \mathbf{z}) + d_{\mathcal{Y}}(\mathbf{x}, \mathbf{z}) \tag{25}$$

(Triangle Inequality)

These spaces have additional flexibility compared to standard metric spaces: note that while $d(y, y) = 0$, $d(x, y) = 0$ does not imply that x and y are identical. The downside of using such spaces, however, is that conventional algebra may not produce the usual results. For example, limits where the distance between a sequence of points and a particular point tends to zero do not convey the same information as in standard metric spaces. However, these odd properties do not concern us, as we only use the spaces for representing a set of distances from our given metric space.

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

F.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 $\mathbf{y} = [y_1, \dots y_d], \mathbf{z} = [z_1, \dots z_d] \in V$,

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

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,*

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

Embedding Algorithms The tool that ensures we can produce isometric embeddings is the following result:

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 embeddable in $\mathbb{R}^{d^+,d^-}$ if and only if,*

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

This bilinear form is very similar to the one used for MDS embeddings [KW78]—it is closely related to the squared distance matrix. The main information needed is what the signature (i.e., how many positive, negative, and zero eigenvalues) of this bilinear form is. If the dimension of the pseudo-Euclidean space we choose to embed in is at least as large as the number of positive and negative eigenvalues, we can obtain isometric embeddings. Because we are working with finite metric spaces, this number is always finite, and, in fact, is never larger than the size of the metric space. This means we can always produce isometric embeddings.

The practical aspects of how to produce the embedding are shown in the first half of Algorithm 1. The basic idea is to do an eigendecomposition and capture eigenvectors corresponding to the positive and negative eigenvalues. These allow us to perfectly reproduce the positive and negative components of the distances separately; the resulting distance is the difference between the two components. The process of performing the eigendecomposition is standard, so that the overall procedure has the same complexity as running MDS. Compare this to MDS: there, we only capture the eigenvectors corresponding to the positive eigenvalues and ignore the negative ones. Otherwise the procedure is identical.

We note that in fact it is possible to embed pseudo-metric spaces isometrically into pseudo-Euclidean spaces, but we never use this fact. Our only application of this tool is to embed conventional metric spaces. However, our results directly lift to this more general setting.

The idea of using pseudo-Euclidean spaces for embeddings that can then be used in kernel-based or other classifiers or other approaches to machine learning is not new. For example, [PPD01] used these spaces for kernel-based learning, [LRBM06] used them for generic pairwise learning, and [PHD$^+$06] showed that they are among several non-standard spaces that provide high-quality representations. Our contribution is using these in the context of weak supervision and learning latent variable models.

Dimensionality We also give more detail on the example we provided showing that pseudo-Euclidean embeddings can have arbitrarily better dimensionality compared to one-hot encodings. The idea here is simple. We start with a particular kind of tree with a root and three branches that are simply long chains (paths) and have t nodes each, for a total of $3t + 1$ nodes. One-hot encodings have dimension that scales with the number of nodes, i.e., dimension $3t + 1$.

Pseudo-euclidean embeddings enable us to embed such a tree into a space of finite (and in fact, very small) dimension while preserving the shortest-hops distances between each pair of nodes in the graph. As described above, the key question is what the number of positive and negative eigenvalues for the squared distance matrix (and thus the bilinear form) is. Fortunately, for such graphs, the signature of the squared-distance matrix is known (Theorem 20 in [BS16]). Applying this result shows that the pseudo-Euclidean dimension is just 3, a tiny fixed value regardless of the value of t above.