A Pretraining Strategies

The loss function (10) is highly non-convex with respect to θ, a consequence of both the objective itself and the nature of hyperbolic neural networks (Ganea et al., 2018b). As a result, we found that initialization plays a crucial role in this problem, since it is very hard to overcome a poor initial local minimum. Even layer-wise random initialization of weights and biases proved futile. As a solution, we experimented with the following three pre-training initialization schemes, all of which intuitively try to approximately ensure (in different ways) that f does not "collapse" the space $\mathcal{Y}$:

Identity. Initialize f_θ to approximate the identity:

$$\min_\theta \sum_{i=1}^{n} d_{\mathbb{D}}(\mathbf{y}_i, f_\theta(\mathbf{y}_i)),$$

which trivially ensures that f_θ (approximately) preserves the overall geometry of the space.

CrossMap. Initialize f_θ to approximately match the target points to the source points in a random permuted order:

$$\min_\theta \sum_{i=1}^{n} d_{\mathbb{D}}(\mathbf{x}_{\sigma(i)}, f_\theta(\mathbf{y}_i))$$

for some permutation $\sigma(i)$, which again ensures that f_θ approximately preserves the global geometry, albeit for an arbitrary labeling of the points.

Procrustes. Following (Bunne et al., 2019), we initialize f_θ to be approximately end-to-end orthogonal:

$$\min_\theta \sum_{i=1}^{n} d_{\mathbb{D}}(f(\mathbf{y}_i), \mathbf{P}\mathbf{y}_i),$$

where $\mathbf{P} = \operatorname{argmin}_{\mathbf{P}\in O(n)} \|\mathbf{X} - \mathbf{P}\mathbf{Y}\|_2^2$, i.e., $\mathbf{P}$ is the solution of (a hyperbolic version of) the Orthogonal Procrustes problem for mapping $\mathbf{Y}$ to $\mathbf{X}$, which can be obtained via singular value decomposition (SVD). This strategy thus requires computing an SVD for every gradient update on θ; hence, it is significantly more computationally expensive than the other two.

B Optimization Details

Each forward pass of the loss function (10) requires solving three regularized OT problems. While this can be done to completion in $O(N^2 \log N\varepsilon^{-3})$ time (Altschuler et al., 2017), practical implementations often run the Sinkhorn algorithm for a fixed number of iterations with a tolerance threshold on the objective improvement. We rely on the `geomloss`[3] package for efficient differentiable Sinkhorn divergence implementation and on the `geoopt`[4] package for Riemannian optimization. We run our method for a fixed number of outer iterations (200 in all our experiments), which given the decay strategy on the entropy regularization parameter ε, ensures that ε ranges from 1×10^1 to 1×10^{-2}. All experiments where run a single machine with 32-core processor, Intel Xeon CPU @3.20 GHz, and exploiting computations on the GPU (a single GeForce Titan X) whenever possible. With this configuration the total runtime of our method on the experiments ranged from < 1 to 20 minutes.

C Dataset Details

To generate the parallel WordNet datasets, we use the `nltk` interface to WordNet, and proceed as follows. In the English WordNet, we first filter out all words except nouns, and generate their transitive closure. For each of the remaining synsets, we query for lemmas in each of the four other languages (Es, Fr, It, Ca), for which `nltk` provides multilingual support in WordNet. These tuples of lemmas form our ground-truth translations, which are eventually split into a validation set of size 5000, leaving all the other pairs for test data (approximately 1500 for each language pairs). Note that the validation is for visualization purposes only, and all model selection is done in a purely unsupervised way based on the training objective. After the multi-lingual synset vocabularies have been extracted, we ensure their transitive closures are complete and write all the relations in these closures to a file, which will be used as an input to the PoincareEmbeddings toolkit.[5]

To generate the datasets for the synthetic noise-sensitivity experiments (§7.2), we start from the original CS-PhD dataset.[6] Given a pre-defined value ν, we iterate through the hierarchy removing node x with probability p, connecting x's children with x's parent to keep the tree connected. We repeat this with noise values $p \in \mathfrak{P} = [0.01, 0.05, 0.1, 0.2]$ and embed all of these using the PoincareEmbeddings in hyperbolic spaces of dimensions $d \in \mathfrak{D} = [2, 5, 10, 20]$. For a given dimensionality and noise level, we use our method to find correspondences between the noise-less and noisy version of the hierarchy (i.e., $|\mathfrak{P}| \times |\mathfrak{N}|$ matching tasks in total).

Statistics about all the datasets used in this work are provided in Table 3. Further details about the OAEI datasets can found on the project's website.[7]

[3] https://www.kernel-operations.io/geomloss/

[4] https://geoopt.readthedocs.io/en/latest/

[5] https://github.com/facebookresearch/poincare-embeddings

[6] http://networkrepository.com/CSphd.php

[7] http://oaei.ontologymatching.org/2018/

	WordNet				Anatomy		Biodiv			
	English (En)	Spanish (Es)	French (Fr)	Catalan (Ca)	Human	Mouse	Flopo	Pto	Envo	Sweet
Entities	8206	8206	8206	8206	3298	2737	360	1456	6461	4365
Relations	47938	47938	47938	47938	18556	7364	472	11283	73881	30101
Embedding Size	10	10	10	10	10	10	10	10	10	10

Table 3: Dataset characteristics. All datasets embedded with the method of Nickel and Kiela (2017).

Model	Metric	Cost	Pretrain	ε-annealing	Layers	Hidden dim	Layer Type	Nonlin.	Opt	LR
Full	Poincare	$d(x,y)$	CrossMap	$10^1 \to 10^{-2}$	10	20	HyperLinear	elu	Radam	10^{-3}
Small	–	–	–	–	2	10	–	–	–	–
Euclidean	Euclidean	–	–	–	–	–	–	–	–	–
ReLU	–	–	–	–	–	–	ReLU	–	–	
Rsgd	–	–	–	–	–	–	–	–	Rsgd	5×10^{-2}
Möbius	–	–	–	–	–	10	Möbius	–	–	–
cosh Cost	–	$-\cosh \circ d_{\mathbb{D}}$	–	–	–	–	–	–	–	–
No Pretrain	–	–	None	–	–	–	–	–	–	–

Table 4: Ablated model configurations for the monolingual En →En WordNet task.

D Model Configurations and Hyperparameters

In Table 4, we provide full configuration details for all the ablated models used in the WordNet En →En self-recovery experiment (results shown in Table 1a). Dashed lines indicate a parameter being the same as in the Full Model.

E A Brief Summary of Theoretical Guarantees for Optimal Transport (Euclidean Case)

As mentioned in Section 4.2, whenever optimal transport is used with the goal of obtaining correspondences, there are various theoretical considerations that become particularly appealing.

The first of such considerations pertains to the nature of the solution, i.e., the optimal coupling π^* which minimizes the cost (5). When the final end goal is to transport points from one space to the other, the best case scenario would be if the optimal π happens to be a "hard" deterministic mapping. A celebrated result by Brenier (1987) (see also (Brenier, 1991)) shows that this indeed the case for the quadratic cost,[8] i.e., for the 2-Wasserstein distance. Even when solving the problem approximately with entropic regularization (cf. Eq. (7)), this result guarantees that the solution found in this way converges to a deterministic mapping as $\varepsilon \to 0$.

Now, assuming now that such a map exists, the next aspect we might be interested in is its smoothness. Intuitively, smoothness of this mapping is desirable since it is more likely to lead to robust matchings in the context of correspondences, even if, again, the argument holds asymptotically for the regularized problem. This, clearly, is a very strong property to require. While not even continuity can be guaranteed in general (Ambrosio and Gigli, 2013), again for the quadratic-cost things are simpler: if the source and target densities are smooth and the support of the target distribution satisfies suitable convexity assumptions, the optimal map is guaranteed to be smooth too (Caffarelli, 1992a; Caffarelli, 1992b).

[8] This result holds in more general settings. We refer the reader to (Santambrogio, 2010; Ambrosio and Gigli, 2013) for further details.

F A Brief Summary of Theoretical Guarantees for Optimal Transport (Riemannian Manifold Case)

Extending the problem beyond Euclidean to more general spaces has been one of the central questions theoretical optimal transport research over the past decades (Villani, 2008). For obvious reasons, here we focus the discussion on results related to hyperbolic spaces, and more generally, to Riemannian manifolds.

Let us first note that Problem (5) is well-defined for any complete and separable metric space $\mathcal{X}$. Since the arclength metric of a Riemannian manifold allows for the direct construction of an accompanying metric space $(\mathcal{X}, d_{\mathcal{X}})$, then OT can be defined over those too. However, some of the theoretical results of their Euclidean counterparts do not transfer that easily to the Riemannian case (Ambrosio and Gigli, 2013). Nevertheless, the existence and uniqueness of the optimal transportation plan π^*, which in addition is induced by

a transport map T, can be guaranteed with mild regularity conditions on the source distribution α. This was first shown in seminal work by McCann (2001). The result, which acts as an Riemannian analogue of that of Brenier for the Euclidean setting (Brenier, 1987), is shown below as presented by Ambrosio and Gigli (2013):

Theorem F.1 (McCann, version of (Ambrosio and Gigli, 2013))**.** *Let M be a smooth, compact Riemannian manifold without boundary and $\alpha \in \mathcal{P}(\mathcal{M})$. Then the following are equivalent:*

(i) $\forall \beta \in \mathcal{P}(\mathcal{M})$, there exists a unique optimal $\pi \in \Pi(\alpha, \beta)$, and this plan is induced by a map T.

(ii) α is regular.

If either (i) or (ii) holds, the optimal T can be written as $x \mapsto \exp_x(-\nabla\phi(x))$ for some c-concave function $\phi : \mathcal{M} \to \mathbb{R}$.

The question of regularity of the optimal map, on the other hand, is much more delicate now than in the Euclidean case (Ambrosio and Gigli, 2013; Ma et al., 2005; Loeper, 2009). In addition to the suitable convexity assumptions on the support of the target density, a restrictive structural condition, known as the Ma-Trudinger-Wang (MTW) condition (Ma et al., 2005), needs to be imposed on the cost in order to guarantee continuity of the optimal map. Unfortunately for our setting, in the case of Riemannian manifolds the MTW condition for the usual quadratic cost $c = d^2/2$ is so restrictive that it implies that $\mathcal{X}$ has non-negative sectional curvature (Loeper, 2009), which rules out hyperbolic spaces. However, a recent sequence of remarkable results (Lee and Li, 2012; Li, 2009) prove that for simple variations of the Riemannian metric d on hyperbolic spaces, smoothness is again guaranteed:

Theorem F.2 (Lee and Li, (Lee and Li, 2012))**.** *Let d be the Riemannian distance function on a manifold of constant sectional curvature -1; then the cost functions $-\cosh \circ d$ and $-\log \circ (1 + \cosh) \circ d$ satisfy the strong MTW condition, and the cost functions $\pm \log \circ \cosh \circ d$ satisfy the weak MTW condition.*

Thus, these cost objectives can be used it out hyperbolic optimal transport matching setting with the hopes of obtaining a smoother solution, and therefore a more stable set of correspondences.