On the Power and Limits of Distance-Based Learning

Periklis A. Papakonstantinou PERIKLIS.RESEARCH@GMAIL.COM
MSIS, Business School, Rutgers University, Piscataway, NJ 08853, USA

Jia Xu JIA.XU@HUNTER.CUNY.EDU
Department of Computer Science, Hunter College, CUNY, 695 Park Ave, New York, NY 10065, USA
& The Graduate Center, CUNY, 365 5th Ave, New York, NY 10016, USA

Guang Yang GUANG.RESEARCH@GMAIL.COM
Institute of Computing Technology, Chinese Academy of Sciences, Beijing, 100190, China
& Aarhus University, Aaogade 34, DK 8200 Aarhus, Denmark

Abstract

We initiate the study of low-distortion finite metric embeddings in multi-class (and multi-label) classification where (i) both the space of input instances and the space of output classes have combinatorial metric structure, and (ii) the concepts we wish to learn are low-distortion embeddings. We develop new geometric techniques and prove strong learning lower bounds. These provable limits hold even when we allow learners and classifiers to get advice by one or more experts. Our study overwhelmingly indicates that post-geometry assumptions are necessary in multi-class classification, as in natural language processing (NLP). Technically, the mathematical tools we developed in this work could be of independent interest to NLP. To the best of our knowledge, this is the first work which formally studies classification problems in combinatorial spaces and where the concepts are low-distortion embeddings.

1. Introduction

Multi-class and multi-label classification, especially when the number of classes is very large, finds important applications in natural language processing, speech recognition, and image recognition. In machine translation, for example, every translated sentence becomes a distinct class. The mathematics of such multi-class settings are far less understood compared to binary or moderate-class-size classification. We put things in new context by introducing metric structure both in the instance space and in the output classes (label space), and we consider the *geometry of the map (concept)* between them. We ask how to effectively combine advice provided by experts in such multi-class tasks.

This is a particularly natural question for multi-class problems. Common, popular machine learning methods — SVMs, kNN, and their variants — are distance-only based. Furthermore, the fact that we have many classes, high-dimensions in the output, and metric structure allows us to study interesting geometric properties (contrast this with binary or low-dimensional spaces of classes).

Can strong geometry help? Let us begin with the following motivating example. Consider the cube, $C_3 = \{0,1\}^3$, equipped with the Hamming distance – e.g. the distance $d_H((0,0,1),(1,0,0)) = 2$ measures the number of positions the two vectors differ. Now, consider the collection of concepts (concept class) $\mathfrak{C} = \{\mathsf{r} : C_3 \to C_3\}$ such that every $\mathsf{r} \in \mathfrak{C}$ preserves the distances between every two strings $x, y \in C_3$, i.e. $d_H(x,y) = d_H(\mathsf{r}(x), \mathsf{r}(y))$. Then, with the strong geometric property (i.e. that distances are preserved) r can be evaluated on every input by knowing $\mathsf{r}(0,0,0), \mathsf{r}(0,0,1), \mathsf{r}(0,1,0), \mathsf{r}(1,0,0)$, since $\mathsf{r}(x_1,x_2,x_3) = x_1 \cdot \mathsf{r}(1,0,0) + x_2 \cdot \mathsf{r}(0,1,0) + x_3 \cdot \mathsf{r}(0,0,1) + (1 - x_1 - x_2 - x_3) \cdot \mathsf{r}(0,0,0)$, with all operations in $\mathsf{GF}(2)$. More generally, if the cube is of size N, we can fully describe r with $1 + \log_2 N$ queries.

In the above example, learning any target concept r is very efficient despite the high dimension of the space. This is because r is an *isometry* (preserves distances exactly),

⋆ **Author names alphabetically ordered** (typical in mathematical literature).

Proceedings of the 33^{rd} International Conference on Machine Learning, New York, NY, USA, 2016. JMLR: W&CP volume 48.

i.e. there is strong underlying geometry. Thus, the example is compatible with the *no-free-lunch theorem* (Wolpert, 1996). It seems, therefore, plausible to consider what happens when we wish to learn a concept with slightly distorted distances – here we consider bi-Lipschitz (formally defined in Section 2) instead of 1-Lipschitz maps since the latter in general is not learnable. Given such a "geometric promise" it is conceivable that new learning methods could be obtained, whereas at the same time lower bounds are more challenging due to the added geometry.

In real-world problems, although distances are strongly preserved this happens only approximately (see the full version). For example, consider the two English sentences "the cat sat on the mat" and "the dog sat on the mat" and their translations to Chinese. These English sentences have *edit distance* exactly 1; edit distance (Levenshtein, 1966) is the smallest number of substitutions/insertions/deletions needed to turn one sentence into the other. We can imagine, even without knowing Chinese, that the Chinese translations also have very small edit distance. In particular, it is smaller than the distance of each of those two translations to the translation of "four centuries and twenty years ago". In other words, under edit distance, English and Chinese become two metric spaces, and (after filtering out some small parts) English-Chinese translation turns into a *low-distortion embedding*. This means that edit distances in the two languages are approximately preserved under human translation. The difficulty of learning concepts that are low-distortion embeddings is the topic this paper studies.

Our classification setting. The set $\mathcal{X}$ of instances and the set $\mathcal{Y}$ of classes (labels) become metric spaces under distance functions $d_{\mathcal{X}}$ and $d_{\mathcal{Y}}$. The goal is to learn an unknown target concept r chosen from a known concept class $\mathfrak{C}$ consisting of low-distortion injective concepts $\mathcal{X} \hookrightarrow \mathcal{Y}$. The above metric spaces are combinatorial in nature, e.g. we consider Hamming and edit (Levenhstein) spaces. These spaces also have very low-diameter induced by strong "geometric connectivity" (aka high-expansion). Roughly speaking, this means that close to any set of points in the space there are many other close-by points (this property makes lower bounds constructions challenging). Finally, we give access to any fixed number of experts $\mathsf{A}_0, \mathsf{A}_1, \mathsf{A}_2, \cdots : \mathcal{X} \to \mathcal{Y}$ who themselves are low-distortion embeddings. Thus, on one hand, this is weaker than the perfect (no distortion) example we gave in the beginning and, on the other hand, stronger since we allow expert advice.

Remark (experiments justifying the setting): This work is theoretical. It deals with general multi-class and low-distortion learning questions. One motivation (in fact, our original question) comes from NLP. In the full version we list empirical results about the low-distortion of human translation maps and expert maps (Google and Bing translation engines).

Given such strong geometric guarantees, is it possible to combine (i.e. a task simpler than general learning) two or more experts in a way that we get close to the target? Our study yields the following conceptual message.

> *In multi-class tasks, such as machine translation and speech recognition, just knowing distances that quantify dissimilarity in the inputs and outputs does not suffice for building or even combining systems, even in the presence of strong underlying geometry.* Thus, "post-geometry structure" is required; *e.g. language models and other constructs that make statistical assumptions about the structure of the correct output.*

Despite the above message, our work is still theoretical. It addresses general questions about learning low-distortion concepts. It is just one reading of our results that rigorously formalizes and proves what before was only intuitively discussed in Statistical NLP (it is common practice in machine learning to formalize intuitive/high-level concepts and principles, as in e.g. the case of Occam's razor).

Note that edit distance has been widely used in computational linguistics, e.g. WER (Word Error Rate is normalized edit distance), to describe sentence similarity. See also in the full version for further experiments on edit distance in natural languages.

Low-distortion regression tasks and on-line experts Before this work, similar problems were studied in the different and more restricted sense of regression. In regression, the problem is to learn a map from, say $\mathbb{R}^n$ to $\mathbb{R}$, with low distortion under a metric used as the loss function. There is a spade of important works in this topic, see e.g. the seminal paper (Alon et al., 1997). Note however that the notion of geometry we care about in this work (inherent in all technical developments of ours) becomes relevant in high-dimensional combinatorial spaces; cf. Chapter 13 of (Matoušek, 2002) for such metric embeddings. Hence, given the structure of spaces and their dimensions, our developments relate to the notion of structured prediction that we discuss below. Finally, note that such regression questions have also been studied in the on-line setting where one gets expert advice; cf. Chapter 11 of (Cesa-Bianchi & Lugosi, 2006) for the most comprehensive to date and in-depth treatment of the literature. This expert advice is on-line, which means that it is not the same as the experts we consider. Our learners have off-line access to the experts. This means that it is (potentially) easier to learn and at the same time lower bounds in our setting are stronger.

Structure in Multi-Class Classification In multi-class classification, very little can be analytically done when there is no structure. Generic existing techniques reduce the problem to binary classification, as in e.g. (Allwein et al., 2001; Rifkin & Klautau, 2004), and are very inefficient. The typical approach in previous works was to exploit structure in $\mathcal{X}$ or in $\mathcal{Y}$ alone. Formalizing the notion of "structure" has received attention in a number of works as in e.g. (Chapelle et al., 2009; Taskar et al., 2004; 2009; Tsochantaridis et al., 2005) within the important framework of structured prediction (Schölkopf et al., 2007). There, structure corresponds to the combinatorial structure of the outputs, identified by trees, lattices, and other forms of combinatorial hierarchical schemes; with the goal to find an embedding $\mathcal{X} \to \mathcal{Y}$ that has low loss on the structured/combinatorial $\mathcal{Y}$. Contrast this to learning a low-distortion embedding; i.e. when *both* spaces are metric spaces and *related through a map* r, where $d_{\mathcal{X}}(x, x')$ is roughly the same (undistorted) as $d_{\mathcal{Y}}(\mathsf{r}(x), \mathsf{r}(x'))$. To make the difference more apparent, think of a structured prediction problem where very similar instances in $\mathcal{X}$ have very different labels in $\mathcal{Y}$; i.e. no low risk classifier for this problem can be a low-distortion embedding. For the same reason, metric labelling (Kleinberg & Tardos, 2002) is different than learning low-distortion embeddings. In other works, "structure" refers to e.g. sparsity properties of the input instances (Hsu et al., 2009) where the number of features/dimensions can be orders of magnitude bigger than in our setting. Finally, there is an important line of work in learning the structure itself. For example,"metric learning" (not to be confused with our work) aims to learn the metric underlying the instance space (Davis & Dhillon, 2008; Jain et al., 2012; Lu et al., 2009). Our framework is very different than all of the above. We consider a new type of structure not merely on $\mathcal{X}$ or $\mathcal{Y}$, but structure on their *metric relation* under a concept $\mathsf{r} : \mathcal{X} \to \mathcal{Y}$ connecting them.

Our Contribution Our technical contribution regards two types of theorems – each stronger than the other in different aspects. Here are details necessary to read our theorems.

We give the learner (learning method) *query access* to (i) the concept r (this gives us the training set) for which we are constructing the classifier, and (ii) two experts $\mathsf{A}_0, \mathsf{A}_1$ (below we explain why the general case reduces to two experts). Query access to a function $f : \mathcal{X} \to \mathcal{Y}$ means that by querying x we only get a label $y = f(x)$ and nothing about an implementation of f. Can the loss of the learner be reduced if, instead of query access, it knows the entire description of $\mathsf{A}_0, \mathsf{A}_1$? The full description of $\mathsf{A}_0, \mathsf{A}_1$ that provide a random labeling is useless ($\mathsf{A}_0, \mathsf{A}_1$ are not related to r). A more interesting example would have been one where for every instance x either $\mathsf{A}_0(x)$ or $\mathsf{A}_1(x)$ is equal to the correct label $\mathsf{r}(x)$, but still knowing everything about $\mathsf{A}_0, \mathsf{A}_1$ cannot help classification. It turns out that such examples exist, though they are non-trivial to construct, and our proofs provide such details.

In our framework, both the learner and the classifier make queries to experts $\mathsf{A}_0, \mathsf{A}_1$. The learner queries $\mathsf{A}_0, \mathsf{A}_1$ when specifying the classifier, and the classifier is given access to $\mathsf{A}_0, \mathsf{A}_1$ when making decisions given an input instance. We distinguish between two types of learners and classifiers: (i) learners/classifiers that know the full description of $\mathsf{A}_0, \mathsf{A}_1$ and (ii) learners/classifiers that make only (polynomially many) queries to $\mathsf{A}_0, \mathsf{A}_1$. Note that the latter type, with only partial knowledge of $\mathsf{A}_0, \mathsf{A}_1$, cannot be stronger than the first one.

Our first theorem is a worst-case impossibility result, which holds true even if the full description of the experts is given to the learner. The second theorem holds for every concept in the class, when the learner only has query access to the experts. In what follows, a *selector* is a classifier that is allowed to make polynomially many queries to two experts, and an *optimal selector* always outputs the best of the labels provided by the two experts.

Theorem 1 (informally stated). *There is a concept class $\mathfrak{C}$, and two experts $\mathsf{A}_0, \mathsf{A}_1 : \mathcal{X} \to \mathcal{Y}$, for edit spaces $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \mathsf{ed})$, such that every learner which (i) knows the full description of $\mathsf{A}_0, \mathsf{A}_1$ and (ii) uses expert advice and polynomially many queries to the target, constructs a selector that has significantly higher loss compared to the optimal selector, for some (worst-case) target concept $\mathsf{r} \in \mathfrak{C}$. Furthermore, A_0, A_1, and r are low-distortion embeddings.*

Theorem 2 (informally stated). *There is a concept class $\mathfrak{C}$ with the same metric properties as in Theorem 1, such that for every target concept r, there exists a set $\mathcal{A}$ of low-distortion experts with the following property. For every selector C constructed by a learner which makes polynomially many queries both to (i) two experts and to (ii) the target concept r, there exist experts $\mathsf{A}_0, \mathsf{A}_1 \in \mathcal{A}$ under which C has unbounded expected loss, while the optimal selector has 0 error.*

The Number of Experts and the Restricted Input Length The above lower bounds would be stronger if we had more than two experts. However, these bounds are not weaker since we can also assume the existence of any fixed number of "dummy" experts whose advice is useless for classification. Note that we consider the case of a fixed input length n, whereas in natural language processing applications n varies. We note that a fixed n only makes the lower bound stronger – i.e. we show that algorithms fail even for a fixed n .

These two theorems are impossibility results. However, a

number of positive technical developments were discovered in the process, and they are of independent interest to theory and applications in machine learning. These include some probabilistic, local finite metric embedding of the edit space to ℓ_p, which can be of independent interest to natural language processing.

At a conceptual level, we initiate the study of low-distortion metric embeddings between "combinatorial" metric spaces. Our notion of embedding is not the same as the informal use of the term in e.g. natural language processing (note that for us, two metric spaces are necessary just to be able to define an embedding). Let us further remark that the embeddings of the edit distance into ℓ_1 or ℓ_2, as one may consider for simplification, face inherent limitations (universal and regardless of dimension of the host space) of non-embeddability (Krauthgamer & Rabani, 2009), whereas the best algorithmic embeddings (Ostrovsky & Rabani, 2007) are even worse. Thus, proving an impossibility result directly on combinatorial metric spaces is much stronger because there is no "information loss" incurred by first moving (extracting features) to $\mathbb{R}^n$.

2. Preliminaries and notation

Algorithms and Resources In this work, lower bounds are information theoretic (stronger than computational lower bounds). Instead of time steps, we measure the number of queries made to various oracles, e.g. experts and target concepts.

Finite Metric Embeddings We consider *spaces* $\mathcal{X}, \mathcal{Y}$ which are sets of size N (exponential in n) where the description of each element is of length n. A *metric space* is denoted by $(\mathcal{X}, d)$, for a space $\mathcal{X}$ and a *distance function* $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The distance function is non-negative, symmetric, and obeys the triangular inequality, i.e. $d(x, y) \geq 0$, $d(x, y) = d(y, x)$ and $d(x, y) \leq d(x, z) + d(z, y)$, for every $x, y, z \in \mathcal{X}$. For $x \in \mathcal{X}$, we denote by $\mathcal{B}_x(\rho) \stackrel{\text{def}}{=} \{y \in \mathcal{X} \mid d(x, y) \leq \rho\}$ the *closed ball of radius ρ centered at x*. The *edit distance* $\mathsf{ed}(x, y)$ makes the set of strings $\{0, 1\}^*$ (or a bigger vocabulary V) a metric space and is defined for $x, y \in \{0, 1\}^*$ as the minimum number of insertions, deletions, and symbol substitutions to edit x into y. The other important metric to this paper is ℓ_1 defined through the norm $\|\cdot\|_1$, where $\|x - y\|_1 \stackrel{\text{def}}{=} \sum_{i=1}^{n} |x_i - y_i|$, for every $x, y \in \mathbb{R}^n$. We informally refer to *edit* and to *Hamming* distance (i.e. ℓ_1 restricted on the cube $\{0, 1\}^n$) as *combinatorial metrics*, to distinguish from other real metric spaces that typically appear in learning literature. Note that edit distance is always a lower bound for Hamming distance, but not the other way around since there are simple example strings with large Hamming but very small edit distance.

For two metric spaces $(\mathcal{X}, d_{\mathcal{X}})$ and $(\mathcal{Y}, d_{\mathcal{Y}})$, an *embedding* of $\mathcal{X}$ into $\mathcal{Y}$ refers to an injective mapping $\phi : \mathcal{X} \to \mathcal{Y}$. The Lipschitz constant (expansion) of ϕ is defined as $\|\phi\|_{\mathsf{Lip}} \stackrel{\text{def}}{=} \sup_{a \neq b} \frac{d_{\mathcal{Y}}(\phi(a), \phi(b))}{d_{\mathcal{X}}(a, b)}$, and the *distortion* of ϕ is defined as $\mathsf{dist}(\phi) \stackrel{\text{def}}{=} \|\phi\|_{\mathsf{Lip}} \times \|\phi^{-1}\|_{\mathsf{Lip}}$. For families of metric spaces $\{(\mathcal{X}_n, \mathcal{Y}_n)\}_{n \geq 1}$ with size parameterized by n and a family of embeddings $\{\phi_n : \mathcal{X}_n \to \mathcal{Y}_n\}_{n \geq 1}$, we say that ϕ is low-distortion if $\mathsf{dist}(\phi)$ is constant with respect to n, and ϕ is *bi-Lipschitz* if both $\|\phi\|_{\mathsf{Lip}}, \|\phi^{-1}\|_{\mathsf{Lip}}$ are bounded by a constant that we call *bi-Lipschitz constant*.

Our Learning Setting We denote by $\mathcal{X}$ the set of possible *instances*, by $\mathcal{Y}$ the set of possible *labels* (or classes), and by r the true concept mapping $\mathcal{X}$ to $\mathcal{Y}$, where $\mathsf{r} : \mathcal{X} \to \mathcal{Y}$ is also known as the *target concept*. A typical goal in learning is to choose the *classifier* (or hypothesis), which is an embedding $\mathcal{X} \to \mathcal{Y}$ that has small generalization error (see below) compared to the target concept $\mathsf{r} \in \mathfrak{C}$, where $\mathfrak{C}$ denotes the *concept class* that captures some kind of prior knowledge about the problem. In PAC learning (Valiant, 1984), a *learner* is an algorithm guaranteed to work for every concept $\mathsf{r} \in \mathfrak{C}$ and for every distribution D over $\mathcal{X}$; the learner is given independent and identically distributed (iid) $\mathsf{poly}(n)$ many *observations* $(s, \mathsf{r}(s))$ sampled from D and constructs the classifier h. In a stronger model (Angluin, 1988) the learner is allowed to make queries to r, i.e. the learner can get correct labels under r for adaptively chosen instances. We aim to construct h with small *generalization error* $\mathcal{R}$ which, for metric spaces, is defined through a loss function $\mathcal{L}(s) \stackrel{\text{def}}{=} d(\mathsf{r}(s), h(s))$, for every $s \in \mathcal{X}$ and $\mathcal{R}(h) \stackrel{\text{def}}{=} \mathbb{E}_{s \sim \mathsf{D}}[\mathcal{L}(s)]$. In our setting, we consider bi-Lipschitz concepts and classifiers. For every target concept r, we are given two experts A_0 and A_1, which are also bi-Lipschitz classifiers. Thus, instead of a learner that constructs a classifier only from observations/queries to r, we consider a *learner with expert advice* and instead of a classifier a *selector* defined below.

Learner with Expert Advice & Selector Given a target concept r and two experts $\mathsf{A}_0, \mathsf{A}_1$, we generalize the notion of learner by utilizing $\mathsf{A}_0, \mathsf{A}_1$ in addition to the observations $(s, \mathsf{r}(s))$ the learner gets. We distinguish between two types of learners. (i) A *learner with expert advice* is given the full description (all instance/label pairs) of the two experts, then makes polynomially many $q(n) = \mathsf{poly}(n)$ queries to the target concept r, and finally outputs the description of a selector. The *selector* is a classifier that on every given input instance outputs the label[1] either of A_0 or A_1. A selector is called *optimal* if on every input instance it always chooses

[1] In fact, our theorems, i.e. Theorem 1, 2 and 4, are much stronger since we allow the selector not just to select the best but also to arbitrarily combine the results of the labelings of $\mathsf{A}_0, \mathsf{A}_1$.

the better result between A_0 and A_1. Note that if we merely know the full description of experts this does not reveal any information about their quality on the unknown target concept. (ii) A *learner with query access to expert advice* is a restriction of the previous learner where instead of knowing the full description of $\mathsf{A}_0, \mathsf{A}_1$, the learner is given query access making at most $q(n)$ queries to $\mathsf{A}_0, \mathsf{A}_1$.

Definition 3 (learner with expert advice)**.** *Let $\mathsf{r} : \mathcal{X} \to \mathcal{Y}$ be the target concept and D be a distribution on $\mathcal{X}$. A* learner with expert advice *(or* learner *for short) is a computationally unbounded process which is given (i) q-many observations in the form $(s, \mathsf{r}(s))$ where s is sampled by the learner, and (ii) the full description of two experts $\mathsf{A}_0, \mathsf{A}_1 : \mathcal{X} \to \mathcal{Y}$, which usually depend on r and D. The learner then constructs a classifier C. We call C a* selector from q-queries *and write $\mathsf{C}_{\mathsf{A}_0,\mathsf{A}_1}$ when emphasizing its dependence on $\mathsf{A}_0, \mathsf{A}_1$. A* learner with query access to expert advice *is given $\mathsf{poly}(n)$ queries to experts $\mathsf{A}_0, \mathsf{A}_1$, instead of the full description.*

3. Limits in learning low-distortion embeddings

Our main and most general results are about edit spaces, which is also motivated by applications to natural language processing.[2] Here, we present two impossibility results. The first theorem is a strong lower bound that holds true for learners with full knowledge of the two experts. In this case, we show that there is *always* at least one concept from the concept class $\mathfrak{C}$ that *cannot* be learned (formally, this is all we need for a PAC lower bound – suffices to exhibit one concept and one distribution). Our second theorem shows the same thing *for every* concept in $\mathfrak{C}$, but for learners with query access to expert advice.

Theorem 1. *For every positive integer $\rho \leq n$ and every function $q = q(n)$, there is a concept class $\mathfrak{C}$, two experts $\mathsf{A}_0, \mathsf{A}_1 : \mathcal{X} \to \mathcal{Y}$ (same experts for every concept) where $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \mathsf{ed})$, such that the following holds true for $\mathsf{D} = U_{2n}$. For every learner with expert advice, there exists a target concept $\mathsf{r} \in \mathfrak{C}$, such that for the constructed selector C from q-queries we have*

$$\mathcal{R}(\mathsf{C}) > \left(\frac{1}{2} - \frac{(4n+4\rho)^{\rho}\, q(n)}{2^{n+1}}\right) \Omega\left(\sqrt{\rho}\right)$$

while the optimal selector[3] has generalization error at most $\frac{(4n+4\rho)^{\rho} q(n)}{2^n} \cdot \rho$. Furthermore, for the given $\mathsf{A}_0, \mathsf{A}_1$ and for every $\mathsf{r} \in \mathfrak{C}$ the bi-Lipschitz constant is less than 5. In particular, for $\rho \leq \left\lfloor \frac{n - \log q(n)}{2 \log n} - 1 \right\rfloor$, $\mathcal{R}(\mathsf{C}) \geq \Omega\left(\sqrt{\rho}\right)$ whereas the generalization error is bounded by $2^{-\Omega(n)}$.

Theorem 2. *For every $\rho \geq 0$, $q(n) \geq 0$, and target concept r with distortion $\leq c$, there exists a set $\mathcal{A}$ of bi-Lipschitz experts defined over $\mathcal{X} \to \mathcal{Y}$, where $\mathcal{X} = \mathcal{Y} = (\{0,1\}^{2n}, \mathsf{ed})$, such that for input distribution $\mathsf{D} = U_{2n}$ the following holds true. For every learner that makes $q(n)$ queries to r and $q(n)$ queries to the two experts, there are two experts $\mathsf{A}_0, \mathsf{A}_1 \in \mathcal{A}$ with distortion $\leq 5\mathsf{c}$, from which the optimal selector has zero error, whereas the learner constructed C has error*

$$\mathcal{R}(\mathsf{C}) \geq \left(\frac{1}{2} - \frac{(4n+4\rho)^{\rho} q(n)}{2^{n+1}}\right) \Omega\left(\frac{\sqrt{\rho}}{c}\right)$$

In particular, for $\rho \leq \left\lfloor \frac{n - \log q(n)}{3 + \log n} - 1 \right\rfloor$, $\mathcal{R}(\mathsf{C}) \geq \Omega\left(\frac{\sqrt{\rho}}{c}\right)$.

In all of our theorems the input length parameter n is linearly related to the length of each element in $\mathcal{X}, \mathcal{Y}$. Theorem 1 and 2 are stated as tradeoffs, while it remains open for exactly zero generalization error in Theorem 1. For any polynomial in n number of queries, i.e. $q(n) = \mathsf{poly}(n)$, the generalization error $\mathcal{R} \geq \Omega\left(n^{0.499}\right)$ for the constructed selector; i.e. learning in this setting is impossible.

3.1. The Warm-Up Setting

We present a warm-up example that illustrates the difficulty of learning an embedding between "combinatorial" metric spaces. This oversimplified artificial setting puts in context the involved analyses of Hamming and edit spaces. We lower bound the generalization error for learners with expert advice by showing the existence of a bad concept in the concept class. To further reduce clutter, we only deal with the uniform distribution over domain $\mathcal{X}$.

Proof idea: We consider concepts $\mathsf{r} \in \mathfrak{C}$ and experts $\mathsf{A}_0, \mathsf{A}_1$ that are low-distortion embeddings $\mathcal{X} \to \mathcal{Y}$. Each embedding is completely specified through a permutation of points as follows. We think of the elements of the domain $\mathcal{X}$ (resp. the range $\mathcal{Y}$) of size 2^n lying on the $N^2 \times N$ lateral surface of a cylinder, where N is exponentially large in the length parameter n and the $N^2 \times N$ grid on this surface defines distance (the distance of two points on the grid is their distance in the space). Then, low-distortion embeddings $\mathsf{r}, \mathsf{A}_0, \mathsf{A}_1$ are all specified by the combinations of sub-permutations such that for every latitude of the cylinder of $\mathcal{X}$, all the points are injectively mapped to the same latitude on the cylinder of $\mathcal{Y}$. Moreover, each sub-permutation preserves relative order of involved points and hence acts as a rotation specified by an offset, where the offsets for every two consecutive latitudes differ by at most one. Thus, we realize every low-distortion embedding as a deformation of the cylinder. To simplify the setting, we equally partition the cylinder into N parts, where each part is a small cylinder (with an $N \times N$ grid) and we only consider

[2] See the full version for details about NLP. In the full version we also state and show the lower bound for the Hamming cubes, which is a stronger result relying on different techniques.

[3] The bound for the optimal selector involves $q(n)$ since the experts $\mathsf{A}_0, \mathsf{A}_1$ depend on the number of queries as in Definition 3.

two deformations on it – "identity" with all zero offsets and "rotating" with increasing offsets (Figure 1.A). By specifying the alternative choices on every small cylinder, we define experts $\mathsf{A}_0, \mathsf{A}_1$ and concepts in $\mathfrak{C}$ (Figure 1.B). The learner with expert advice cannot work because there are always concepts in $\mathfrak{C}$ that are far away from the target concept $\mathsf{r} \in \mathfrak{C}$ but indistinguishable from r to the learner.

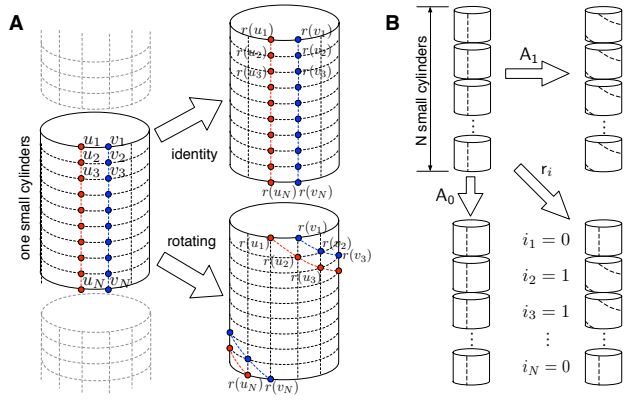

Figure 1. The warm-up example. (A) Two kinds of deformations on a small cylinder: "identity" preserves the position of every element; "rotating" maps each element to a position on the same latitude by applying rotations on every latitude. (B) The two experts $\mathsf{A}_0, \mathsf{A}_1$ and the concept $\mathsf{r}_i \in \mathfrak{C}$, where r_i is a mixture of $\mathsf{A}_0, \mathsf{A}_1$ specified by its subscript $i \in \{0,1\}^N$.

Theorem 4. *There exists a concept class $\mathfrak{C}$ and two experts $\mathsf{A}_0, \mathsf{A}_1 : \mathcal{X} \to \mathcal{Y}$ (same for every concept in the class), where $|\mathcal{X}| = |\mathcal{Y}| = 2^n$, such that when input follows the uniform distribution $\mathsf{D} = U_n$ the following holds true. For every learner that constructs C from q-queries, there exists $\mathsf{r} \in \mathfrak{C}$ such that*

$$\mathcal{R}(\mathsf{C}) \geq (1 - \frac{q}{2^{n/3}}) \max\{\mathcal{R}(\mathsf{A}_0), \mathcal{R}(\mathsf{A}_1)\}$$

Furthermore, the optimal selector makes zero error and all concepts in $\mathfrak{C}$ and $\mathsf{A}_0, \mathsf{A}_1$ are embeddings from $\mathcal{X}$ to $\mathcal{Y}$ with bi-Lipschitz constant ≤ 2.

Thus, for $q(n) = \mathsf{poly}(n)$, every selector has worst-case generalization error asymptotical to that of the worst of the two experts, although an optimal selection between these two experts leads to zero error.

Proof. We first define the concept class $\mathfrak{C}$ and the experts $\mathsf{A}_0, \mathsf{A}_1$ with length parameter n, and then analyze the generalization error of the selector. The bi-Lipschitz conditions follows immediately.

Construction (the warm-up example) The two experts $\mathsf{A}_0, \mathsf{A}_1$ and concepts in $\mathfrak{C}$ are all defined over cylindric metric spaces as follows.

Definition 5. *The* cylindric metric space *is the space $[N^2] \times [N]$ equipped with the following distance function*

$$d\big((x,y),(x',y')\big) = |x - x'| + |(y - y') \bmod N|$$

for all $(x,y),(x',y') \in [N^2] \times [N]$ in the space, where the modular operation $(y - y') \bmod N$ takes values from $\big\{z \in \mathbb{Z} \big| -N/2 < z \leq N/2\big\}$. Thus, $(x,y) = (x,y')$ if and only if $y \equiv y' \bmod N$.

Let instances and labels be drawn from the above cylindric spaces, i.e. $\mathcal{X} = \mathcal{Y} = [N^2] \times [N]$ where $N = 2^{n/3}$ and hence $|\mathcal{X}| = |\mathcal{Y}| = 2^n$. We consider experts $\mathsf{A}_0, \mathsf{A}_1$ defined as follows:

$$\mathsf{A}_0(x,y) \stackrel{\text{def}}{=} (x,y), \qquad \mathsf{A}_1(x,y) \stackrel{\text{def}}{=} \big(x, (y+x) \bmod N\big)$$

Let the concept class be $\mathfrak{C} \stackrel{\text{def}}{=} \{\mathsf{r}_i \big| \ i \in \{0,1\}^N, \|i\|_1 = N/2\}$, where each concept $\mathsf{r}_i : \mathcal{X} \to \mathcal{Y}$ indexed by $i = (i_1, i_2, \cdots, i_N)$ is defined as

$$\mathsf{r}_i(x,y) \stackrel{\text{def}}{=} \big(x, (y + x \times i_{\lceil x/N \rceil}) \bmod N\big) \tag{1}$$

The intuition is that we partition the cylindric space into N small cylinders for $\lceil x/N \rceil = 1, 2, \ldots, N$ respectively, where each small cylinder has size $N \times N$. Formally, for every $j \in [N]$ and $(j-1)N + 1 \leq x \leq jN$, it holds that $\mathsf{r}_i(x,y) = \mathsf{A}_{i_j}(x,y)$. Recalling that the index i satisfies $\|i\|_1 = N/2$, we conclude $\mathcal{R}(\mathsf{A}_0) = \mathcal{R}(\mathsf{A}_1) = \frac{N}{8}$ for every target concept $\mathsf{r}_i \in \mathfrak{C}$.

Now, we show the lower bound for an arbitrary selector constructed by any learner with expert advice that makes q queries to the target concept r_i.

Lower Bound for $\mathcal{R}(\mathsf{C})$ We lower bound the generalization error of the best selector C constructed from q-queries. Since the mapping r_i on every small cylinder is specified by a single bit of the subscript i following (1), the label for every instance is uniquely determined by one bit in i. Without loss of generality, we may assume that the labels for the q many queries are fully determined by the first q bits of i. For the subscript $i \in \{0,1\}^N$, let $i' = (i_1, i_2, \cdots, i_q, 1 - i_{q+1}, 1 - i_{q+2}, \cdots, 1 - i_N)$ be the string that flips the last $N - q$ bits of i. Then, we prove that $\mathcal{R}(\mathsf{C}) \geq \frac{N-q}{8}$ either for target concept being r_i or $\mathsf{r}_{i'}$.

$$\begin{aligned}
&\mathcal{R}_{\mathsf{r}_i}(\mathsf{C}) + \mathcal{R}_{\mathsf{r}_{i'}}(\mathsf{C}) \\
=&\mathbb{E}_{s\sim\mathsf{D}}\Big[d\big(\mathsf{r}_i(s), \mathsf{C}(s)\big)\Big] + \mathbb{E}_{s\sim\mathsf{D}}\Big[d\big(\mathsf{r}_{i'}(s), \mathsf{C}(s)\big)\Big] \\
\geq&\mathbb{E}_{s\sim\mathsf{D}}\Big[d\big(\mathsf{r}_i(s), \mathsf{r}_{i'}(s)\big)\Big] \geq \frac{N-q}{N} \cdot \frac{N}{4} = \frac{N-q}{4}
\end{aligned}$$

Recalling that $N = 2^{n/3}$ and $\mathcal{R}(\mathsf{A}_0) = \mathcal{R}(\mathsf{A}_1) = \frac{N}{8}$,

$$\mathcal{R}(\mathsf{C}) \geq \frac{N-q}{8} = (1 - \frac{q}{2^{n/3}}) \max\{\mathcal{R}(\mathsf{A}_0), \mathcal{R}(\mathsf{A}_1)\}$$

□

3.2. What Changes in Hamming and Edit Spaces?

The warm-up example illustrates the lower bound phenomena, though it relies on the nice properties of the artificial metric structure. In particular, it has dimension two and a large diameter (i.e. exponential in n), as well as a tailored (almost ℓ_1) distance function. However, natural metric spaces lack such properties (the warm-up example was engineered for the lower bound to work). They may have much higher dimensions (i.e. linear in n) and lower diameters (i.e. logarithmic in n), and be equipped with more natural/common distance functions, e.g. ℓ_1 or edit distance ed. Such metric spaces can hardly be visualized and the low-distortion embedding over these spaces do not have an immediate decomposition as in the warm-up example.

Important examples of natural metric spaces are high dimensional spaces (i.e. $\{0,1\}^n$) equipped with Hamming distance or edit distance. The Hamming space $(\{0,1\}^n, \ell_1)$ has diameter n, which is logarithmic to its size 2^n. For $x \in \{0,1\}^n$ and radius $\rho > 0$, the size of $\mathcal{B}_x(\rho)$ is exponential in ρ, compared to what was just polynomial (i.e. $\mathrm{O}\left(\rho^2\right)$) in the warm-up example. In this sense, the Hamming space has much stronger geometric relation among neighborhoods than the warm-up example, and hence the low-distortion property provides more information. Note that edit distance is always upper bounded by Hamming distance, the geometric relation is strengthened in the edit space $(\{0,1\}^n, \mathsf{ed})$.

3.3. Overview of the Edit Space Impossibility

We present a high-level description for the worst-case impossibility (Theorem 1). The full proof and the proof for Hamming spaces is given in the full version.

The intuition of Theorem 1 is that each observation $(s, \mathsf{r}(s))$ only provides information about instances that are close to s in $\mathcal{X}$ (or respectively labels close to $\mathsf{r}(s)$ in $\mathcal{Y}$) following the low-distortion property of r, whereas the distortion grows large enough for far away instances such that labels cannot be deduced. We think of each observation under the low-distortion embedding as a "beacon", which reveals information for instances within a fixed radius ρ. Therefore, the number of points that are influenced by those "beacons" upper bounds the accuracy of a selector from q-queries.

Proof sketch of Theorem 1. For the full argument (including a proof of the bi-Lipschitz condition) see the full version. Here, we first present the constructions of experts $\mathsf{A}_0, \mathsf{A}_1$ and the concept class $\mathfrak{C}$, and then lower bound $\mathcal{R}(\mathsf{C})$ by analyzing the influence of "beacons".

Construction 6. *For $x \in \{0,1\}^n, y \in \{0,1\}^n$, let the experts be as follows:*

$$\mathsf{A}_0(x,y) \stackrel{def}{=} (x,y), \qquad \mathsf{A}_1(x,y) \stackrel{def}{=} \big(x, y + m(\rho)\big) \tag{2}$$

where the computation of $y + m(\rho)$ is bitwise over $\{0,1\}^n$, and we define the masking function $m : \{0,1,\cdots,n\} \to \{0,1\}^n$ that for every $t \in \{0,1,\ldots,n\}$

$$m(t) \stackrel{def}{=} (\underbrace{0,\cdots,0}_{n-t},\underbrace{1,1,\cdots,1}_{t}) \tag{3}$$

Let $\mathfrak{C} \stackrel{def}{=} \{\mathsf{r}_0\} \bigcup \left\{\mathsf{r}_Q \middle| Q \subseteq \{0,1\}^{2n}, |Q| \le q(n)\right\}$, where $\mathsf{r}_0 = \mathsf{A}_0$, and for every Q the concept r_Q is defined as $\mathsf{r}_Q(x,y) \stackrel{def}{=} \Big(x, y + m\big(w_Q(x)\big)\Big)$, for $w_Q(x) \stackrel{def}{=} \min\left\{\min_{q_i' \in Q'}\left\{\mathsf{ed}\big(x, q_i'\big)\right\}, \rho\right\}$ and $Q' \stackrel{def}{=} \left\{q_i' \middle| \exists q_i \in Q, q_i = (q_i', q_i'')\right\}$.

Now, we discuss the edit distance between y and $y + m(\rho)$. By definition, $y + m(\rho)$ refers to the string obtained by flipping the last ρ bits of y, whose effect on resulted edit distance is lower bounded by Lemma 7. This lemma is proved in the full version by calculating the probability that there are at least $(\rho + \sqrt{\rho})/2$ many 0's in the last ρ uniform random bits.

Lemma 7. *For every $\rho \le n$ and m as in (3),*

$$\mathop{\mathbb{E}}_{y \sim U_n}\left[\mathsf{ed}\big(y, y + m(\rho)\big)\right] \ge \Omega\left(\sqrt{\rho}\right)$$

Lemma 7 lower bounds the expected distance between $\mathsf{A}_0(s)$ and $\mathsf{A}_1(s)$ which is also (linearly) related to the generalization error of a selector that randomly selects between A_0 and A_1. However, no better treatment exists for an instance that is far from all observations, since those far away "beacons" reveals no useful information about it. Thus, to lower bound $\mathcal{R}(\mathsf{C})$, we first upper bound the amount of information that can be deduced from all the $q(n)$ observations.

Upper Bound for the Influence of "Beacons". Let Q be the set of all queries instances, i.e. the "beacons". By definition, $\mathsf{r}_Q(x,y) = \mathsf{r}_0(x,y)$ only when $x \in Q'$, and the distance of r_0 and r_Q gradually grows to ρ as the instance moves away from "beacons". Since each beacon influences a ball of radius ρ which has size at most $(4(n+\rho))^\rho$, the influence of all beacons is upper bounded by the union of balls.

$$\left|\bigcup_{q_i' \in Q'} \mathcal{B}_{q_i'}(\rho)\right| < (4n + 4\rho)^\rho \, q(n) \tag{4}$$

Lower Bound for $\mathcal{R}(\mathsf{C})$ By inequality (4) and Lemma 7,

$$\begin{aligned}
&\mathsf{ed}(\mathsf{r}_0,\mathsf{r}_Q) = \mathop{\mathbb{E}}_{(x,y)\sim U_{2n}}\left[\mathsf{ed}\big(\mathsf{r}_0(x,y),\mathsf{r}_Q(x,y)\big)\right]\\
=&\frac{1}{2^{2n}}\sum_{(x,y)\in\{0,1\}^{2n}}\mathsf{ed}\Big((x,y),\big(x,y+m(w_Q(x))\big)\Big)\\
\geq&\frac{1}{2^n}\sum_{\substack{x\in\{0,1\}^n\\ w_Q(x)=\rho}}\frac{1}{2^n}\sum_{y\in\{0,1\}^n}\mathsf{ed}\Big((x,y),\big(x,y+m(w_Q(x))\big)\Big)\\
>&\frac{2^n-(4n+4\rho)^\rho\,q(n)}{2^n}\mathop{\mathbb{E}}_{y\sim U_n}\left[\mathsf{ed}\big(y,y+m(\rho)\big)\right]\\
\geq&\left(1-\frac{(4n+4\rho)^\rho\,q(n)}{2^n}\right)\Omega\left(\sqrt{\rho}\right)
\end{aligned}$$

Since C cannot distinguish r_0 from r_Q, there exists $\mathsf{r}\in\mathfrak{C}$ (in particular, $\mathsf{r}\in\{\mathsf{r}_0,\mathsf{r}_Q\}$) such that

$$\mathcal{R}(\mathsf{C})\geq\mathsf{ed}(\mathsf{r}_0,\mathsf{r}_Q)/2>\left(\frac{1}{2}-\frac{(4n+4\rho)^\rho\,q(n)}{2^{n+1}}\right)\Omega\left(\sqrt{\rho}\right)$$

See the full version for the proof of bi-Lipschitz condition. □

4. Conclusions and Future Work

This work initiates the study of learning low-distortion maps between natural combinatorial metric spaces. If there is no distortion (isometry) then learning becomes very efficient. What we show here is that in general, tiny small distortion turns the problem not learnable. In fact, this is shown in a much stronger sense, where we are given (even off-line) access to experts.[4]

Note that our work just introduces the setting, while the main problem is still how to get positive results, i.e. which additional natural properties would give rise to algorithms. Along this line, we can think of two main research directions in this framework, which are not tackled by our work.

The first direction is to investigate natural geometric/metric conditions (in addition to low-distortion) that suffice to devise learning methods.

The second direction builds on top of the first one. We would like to know whether one can solve the on-line learning version of our problem. Again, further restrictions should be identified. The study of the on-line version is a potentially very fruitful direction.

[4]To prove our theorems we develop new technical tools. Some of them could be of independent interest to NLP (e.g. the randomized algorithm that computes the local embedding, implicit in the proof of Lemma 10 in the full version).

Acknowledgements

We wish to thank the anonymous reviewers for the very useful suggestions and pointers to the literature. G.Y. acknowledges partial support from National Natural Science Foundation of China (61222202, 61433014, 61502449) and the Danish National Research Foundation and The National Science Foundation of China (under the grant 61361136003) for the Sino-Danish Center for the Theory of Interactive Computation and from the Center for Research in Foundations of Electronic Markets (CFEM), supported by the Danish Strategic Research Council.

References

Allwein, E. L, Schapire, R. E, and Singer, Y. Reducing multiclass to binary: A unifying approach for margin classifiers. *The Journal of Machine Learning Research (JMLR)*, 1:113–141, 2001.

Alon, N., Ben-David, S., Cesa-Bianchi, N., and Haussler, D. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4): 615–631, 1997.

Angluin, D. Queries and concept learning. *Machine learning*, 2(4):319–342, 1988.

Cesa-Bianchi, N. and Lugosi, G. *Prediction, learning, and games*. Cambridge university press, 2006.

Chapelle, O., Do, C. B., Teo, C. H., Le, Q. V, and Smola, A. J. Tighter bounds for structured estimation. In *Advances in neural information processing systems*, pp. 281–288, 2009.

Davis, J. V and Dhillon, I. S. Structured metric learning for high dimensional problems. In *SIGKDD*, pp. 195–203. ACM, 2008.

Hsu, D., Kakade, S., Langford, J., and Zhang, T. Multi-label prediction via compressed sensing. In *NIPS*, volume 22, pp. 772–780, 2009.

Jain, P., Kulis, B., Davis, J. V, and Dhillon, I. S. Metric and kernel learning using a linear transformation. *The Journal of Machine Learning Research (JMLR)*, 13(1): 519–547, 2012.

Kleinberg, J. and Tardos, E. Approximation algorithms for classification problems with pairwise relationships: Metric labeling and markov random fields. *Journal of the ACM (JACM)*, 49(5):616–639, 2002.

Krauthgamer, R. and Rabani, Y. Improved lower bounds for embeddings into l_1. *SIAM Journal on Computing*, 38(6):2487–2498, 2009.

Levenshtein, V. I. Binary codes capable of correcting deletions, insertions and reversals. In *Soviet physics doklady*, volume 10, pp. 707, 1966.

Lu, Z., Jain, P., and Dhillon, I. S. Geometry-aware metric learning. In *ICML*, pp. 673–680. ACM, 2009.

Matoušek, J. *Lectures on discrete geometry*, volume 212. Springer, 2002.

Ostrovsky, R. and Rabani, Y. Low distortion embeddings for edit distance. *Journal of the ACM (JACM)*, 54(5):23, 2007.

Rifkin, R. and Klautau, A. In defense of one-vs-all classification. *The Journal of Machine Learning Research (JMLR)*, 5:101–141, 2004.

Schölkopf, B., Smola, A. J, Taskar, B., and Vishwanathan, S.V.N. *Predicting Structured Data*. MIT Press, 2007.

Taskar, B., Guestrin, C., and Koller, D. Max-margin markov networks. *NIPS*, 16:25, 2004.

Taskar, B., Weiss, D., Sapp, B., and Toshev, A. Structured prediction cascades. In *NIPS'09 workshop on approximate learning of large scale graphical models: theory and applications*, 2009.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y. Large margin methods for structured and interdependent output variables. In *The Journal of Machine Learning Research (JMLR)*, pp. 1453–1484, 2005.

Valiant, L. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Wolpert, D. H. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.