Multiple Instance Learning with Manifold Bags

Boris Babenko[1] BBABENKO@CS.UCSD.EDU
Nakul Verma[1] NAVERMA@CS.UCSD.EDU
Piotr Dollár[2] PDOLLAR@CALTECH.EDU
Serge Belongie[1] SJB@CS.UCSD.EDU
[1]University of California, San Diego, CA
[2]California Institute of Technology, Pasadena, CA

Abstract

In many machine learning applications, labeling every instance of data is burdensome. Multiple Instance Learning (MIL), in which training data is provided in the form of labeled bags rather than labeled instances, is one approach for a more relaxed form of supervised learning. Though much progress has been made in analyzing MIL problems, existing work considers bags that have a *finite* number of instances. In this paper we argue that in many applications of MIL (e.g. image, audio, etc.) the bags are better modeled as low dimensional *manifolds* in high dimensional feature space. We show that the geometric structure of such manifold bags affects PAC-learnability. We discuss how a learning algorithm that is designed for finite sized bags can be adapted to learn from manifold bags. Furthermore, we propose a simple heuristic that reduces the memory requirements of such algorithms. Our experiments on real-world data validate our analysis and show that our approach works well.

1. Introduction

Traditional supervised learning requires example/label pairs during training. However, in many domains labeling every single instance of data is either tedious or impossible. The Multiple Instance Learning framework (MIL), introduced by Dietterich et al. (1997), provides a general paradigm for a more relaxed form of supervised learning: instead of receiving example/label pairs, the learner gets unordered sets of instances, or *bags*, and labels are provided for each bag, rather than for each instance. A bag is labeled positive if it contains at least one positive instance. In recent years MIL has received significant attention in terms of both algorithm design and applications (Maron & Lozano-Perez, 1998; Andrews et al., 2002; Zhang & Goldman, 2002; Viola et al., 2005).

Theoretical PAC-style analysis of MIL problems has also seen progress in the last decade (Auer et al., 1997; Blum & Kalai, 1998; Long & Tan, 1998; Sabato & Tishby, 2009; Sabato et al., 2010). Typical analysis formulates the MIL problem as follows: a fixed number of instances, r, is drawn from an instance space $\mathcal{I}$ to form a bag. The sample complexity for bag classification is then analyzed in terms of the bag size (r). Most of the theory work has focused on reducing the dependence on r under various settings. For example, Blum & Kalai (1998) showed that if one has access to a noise tolerant learner and the bags are formed by drawing r independent samples from a fixed distribution over $\mathcal{I}$, then the sample complexity grows linearly with r. Recently, Sabato & Tishby (2009) showed that if one can minimize the empirical error on bags, then even if the instances in a bag have arbitrary statistical dependence, sample complexity grows only logarithmically with r.

The above line of work is rather restrictive. Any dependence on r makes it impossible to apply these generalization bounds to problems where bags have infinitely many instances – a typical case in practice. Consider the following motivating example: we would like to predict whether an image contains a face (as in Viola et al., 2005). Putting this in the MIL framework, a bag is an entire image, which is labeled positive if and only if there is a face in that image. The individual instances are image patches. Notice that in this scenario the instances collectively form (a discrete approximation to) a low-dimensional manifold; see Figure 1. Here we expect the sample complexity to scale with the geometric properties of the underly-

Appearing in *Proceedings of the 28th International Conference on Machine Learning*, Bellevue, WA, USA, 2011.

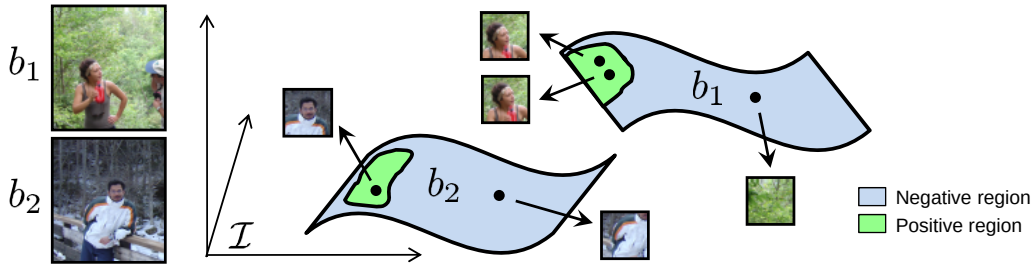


Figure 1. **Manifold Bags:** In this example the task is to predict whether an image contains a face. Each bag is an image, and individual instances are image patches of a fixed size. Examples of two positive bags b_1 and b_2 (left), and a visualization of the instance space $\mathcal{I}$ (right) are shown. The two bags trace out low-dimensional manifolds in $\mathcal{I}$; in this case the manifold dimension is two since there are two degrees of freedom (the x and y location of the image patch). The green regions on the manifolds indicate the portion of the bags that is positive.

ing manifold bag rather than the number of instances per bag.

This situation arises in many other MIL applications where some type of sliding window is used to break up an object into many overlapping pieces: images (Andrews et al., 2002; Viola et al., 2005), video (Ali & Shah, 2008; Buehler et al., 2009), audio (Saul et al., 2001; Mandel & Ellis, 2008), and sensor data (Stikic & Schiele, 2009). Consider also the original molecule classification task that motivated Dietterich et al. (1997) to develop MIL, where a bag corresponds to a molecule, and instances are different shapes that molecule can assume. Even in this application, "as the molecule changes its shape, it traces out a manifold through [feature] space" (Maron & Lozano-Perez, 1998). Thus, manifold structure is an integral aspect of these problems that needs to be taken into account in MIL analysis and algorithm design.

In this work we analyze the MIL framework for bags containing potentially infinite instances. In this setting a bag is drawn from a bag distribution, and is labeled positive if it contains at least one positive instance. In order to have a tractable analysis, we impose a structural constraint on the bags: we assume that bags are low dimensional manifolds in the instance space, as discussed above. We show that the geometric structure of such bags is intimately related to the PAC-learnability of MIL problems. We investigate how learning is affected if we have have access to only a limited number of instances per manifold bag. We then discuss how existing MIL algorithms, that are designed for finite sized bags, can be adapted to learn from manifold bags efficiently using an iterative querying heuristic. Our experiments on real-world data (image and audio) validate the intuition of our analysis and show that our querying heuristic works well in practice.

2. Problem Formulation and Analysis

Let $\mathcal{I}$ be the domain of instances (for the purposes of our discussion we assume it to be $\mathbb{R}^N$ for some large N), and let $\mathcal{B}$ be the domain of bags. Here we impose a structural constraint on $\mathcal{B}$: each bag from $\mathcal{B}$ is a low dimensional manifold over the instances of $\mathcal{I}$. More formally, each bag $b \in \mathcal{B}$ is a smooth bijection[1] from $[0,1]^n$ to some subset of $\mathcal{I}$ ($n \ll N$). The geometric properties of such bags are integral to our analysis. We will thus do a quick review of the various properties of manifolds that will be useful in our discussion.

2.1. Differential Geometry Basics

Let f be a smooth bijective mapping from $[0,1]^n$ to $M \subset \mathbb{R}^N$. We call the image of f (i.e. M) a manifold (note that M is compact and has a boundary). The *dimension* of the domain (n in our case) corresponds to the latent degrees of freedom and is typically referred to as the intrinsic dimension of the manifold M (denoted by $\text{DIM}(M)$).

Since one of the key quantities in classic analysis of MIL is the bag size, we require a similar quantity to characterize the "size" of M. One natural way to characterize this is in terms of the volume of M. The *volume* (denoted by $\text{VOL}(M)$) is given by the quantity $\int_{u_1,\ldots,u_n} \sqrt{\det(J^\intercal J)} du_1 \ldots du_n$, where J is the $N \times n$ Jacobian matrix of the function f, with individual entries defined as $J_{ij} := \partial f_i / \partial u_j$.

Unlike a finite size bag, a finite volume manifold $M \subset \mathbb{R}^N$ can be arbitrarily "complex" – it can twist and turn in all sorts of ways in the surrounding space. We therefore need to also get a handle on its curvi-

[1]Here we are only considering a restricted class of manifolds – those that are globally diffeomorphic to $[0,1]^n$. This is only done for convenience. The results here are generalizable to arbitrary (compact) manifolds.

ness. Borrowing the notation from computational geometry literature, we can characterize the complexity of M via its condition number (see Niyogi et al., 2006). We say that the *condition number* of M (denoted by $\text{COND}(M)$) is $\frac{1}{\tau}$, if τ is the largest number such that the normals of length $r < \tau$ at any two distinct points in M don't intersect. One can bound the sectional curvature of M at any point by $1/\tau$. Hence, when τ is large, the manifold is relatively flat and vice versa.

With these definitions, we can define a structured family of bag spaces.

Definition 1 *We say that a bag space $\mathcal{B}$ belongs to class (V, n, τ), if for every $b \in \mathcal{B}$, we have*[2] *that* $\text{DIM}(b) = n$, $\text{VOL}(b) \leq V$, *and* $\text{COND}(b) \leq 1/\tau$.

In what follows, we will assume that $\mathcal{B}$ belongs to class (V, n, τ). We now provide our main results, with all the supporting proofs in the Appendix.

2.2. Learning with Manifold Bags

Since we are interested in PAC-style analysis, we will be working with a fixed hypothesis class $\mathcal{H}$ over the instance space $\mathcal{I}$ (that is, each $h \in \mathcal{H}$ is of the form $h : \mathcal{I} \to \{0, 1\}$). The corresponding *bag* hypothesis class $\overline{\mathcal{H}}$ over the bag space $\mathcal{B}$ (where each $\bar{h} \in \overline{\mathcal{H}}$ is of the form $\bar{h} : \mathcal{B} \to \{0, 1\}$) is defined as the set of classifiers $\{\bar{h} : h \in \mathcal{H}\}$ where, for any $b \in \mathcal{B}$, $\bar{h}(b) \stackrel{\text{def}}{=} \max_{\alpha \in [0,1]^n} h(b(\alpha))$. We assume that there is some unknown instance classification rule $h^* : \mathcal{I} \to \{0, 1\}$ that gives the true labels for all instances.

The learner gets access to m bag/label pairs $(b_i, y_i)_{i=1}^m$, where each bag b_i is drawn independently from an unknown but fixed distribution $\mathcal{D}_\mathcal{B}$ over $\mathcal{B}$, and is labeled according to the MIL rule $y_i \stackrel{\text{def}}{=} \max_{\alpha \in [0,1]^n} h^*(b_i(\alpha))$. We denote a sample of size m as S_m.

Our learner should ideally return the hypothesis $\bar{h}$ that achieves the lowest bag generalization[3] error: $\text{err}(\bar{h}) \stackrel{\text{def}}{=} \mathbb{E}_{b \sim \mathcal{D}_\mathcal{B}}[\bar{h}(b) \neq y]$. This, of course, is not possible as the learner typically does not have access to the underlying data distribution $\mathcal{D}_\mathcal{B}$. Instead, the learner has access to the sample S_m, and can minimize the *empirical* error: $\widehat{\text{err}}(\bar{h}, S_m) \stackrel{\text{def}}{=} \frac{1}{m} \sum_{i=1}^m \mathbb{1}\{\bar{h}(b_i) \neq y_i\}$. Various PAC results relate these two quantities in terms of the properties of $\overline{\mathcal{H}}$.

[2]Technically b is a function and *not* a manifold. For readability, we will occasionally abuse the notation and use b to mean the manifold produced by the image of b in the instance space $\mathcal{I}$.

[3]One can also talk about the generalization error over instances. As noted in previous work (e.g. Sabato & Tishby, 2009), PAC analysis of the instance error typically requires stronger assumptions.

Perhaps the most obvious way to bound $\text{err}(\bar{h})$ in terms of $\widehat{\text{err}}(\bar{h}, S_m)$ is by analyzing the VC-dimension of the bag hypotheses, $\text{VC}(\overline{\mathcal{H}})$, and applying the standard VC-bounds (see e.g. Vapnik & Chervonenkis, 1971). While finding the VC-dimension of the bag hypothesis class is non-trivial, the VC-dimension of the corresponding *instance* hypotheses, $\text{VC}(\mathcal{H})$, is well known for many popular choices of $\mathcal{H}$. Sabato & Tishby (2009) showed that for finite sized bags the VC-dimension of *bag* hypotheses (and thus the generalization error) can be bounded in terms of the VC-dimension of the underlying *instance* hypotheses. Although one might hope that this analysis could be carried over to bags of infinite size that are well structured, this turns out to not be the case.

2.2.1. $\text{VC}(\overline{\mathcal{H}})$ is Unbounded for Arbitrarily Smooth Manifold Bags

We begin with a surprising result which goes against our intuition that requiring bag smoothness should suffice in bounding $\text{VC}(\overline{\mathcal{H}})$. We demonstrate that requiring the bags to be low-dimensional, arbitrarily flat manifolds with fixed volume is not enough to get a handle on generalization error even for one of the simplest instance hypothesis classes (set of hyperplanes in $\mathbb{R}^N$). In particular,

Theorem 2 *For any $V > 0$, $n \geq 1$, $\tau < \infty$, let $\mathcal{B}$ contain all manifolds M such that $\dim(M) = n$, $\text{VOL}(M) \leq V$, and $\text{COND}(M) \leq 1/\tau$ (i.e. $\mathcal{B}$ is the largest member of class (V, n, τ)). Let $\mathcal{H}$ be the set of hyperplanes in $\mathbb{R}^N$ ($N > n$). Then for any $m \geq 1$, there exists a set of m bags $b_1, \ldots, b_m \in \mathcal{B}$, such that the corresponding bag hypothesis class $\overline{\mathcal{H}}$ (over the bag space $\mathcal{B}$) realizes all possible 2^m labelings.*

Thus, $\text{VC}(\overline{\mathcal{H}})$ is unbounded making PAC-learnability seemingly impossible. To build intuition for this apparent richness of $\overline{\mathcal{H}}$, and possible alternatives to bound the generalization error, let us take a quick look at the case of one-dimensional manifolds in $\mathbb{R}^2$ with halfspaces as our $\mathcal{H}$. For any m, we can place a set of m manifold bags in such a way that all labelings are realizable by $\overline{\mathcal{H}}$ (see Fig. 2 for an example where $m = 3$; see Appendix A.1 for a detailed construction).

The key observation is that in order to label a bag positive, the instance hypothesis needs to label just a *single* instance in that bag positive. Considering that our bags have an infinite number of points, the positive region can occupy an arbitrarily small fraction of a positively labeled bag. This gives our bag hypotheses immense flexibility even when the underlying instance hypotheses are quite simple.

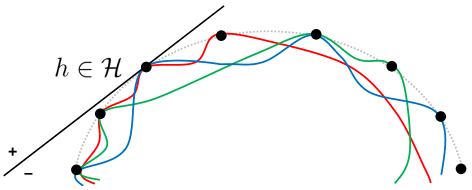
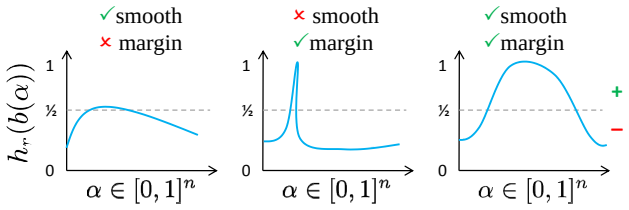


Figure 2. **Bag hypotheses over manifold bags have unbounded VC-dimension.** Three bags (colored blue, green and red) go around the eight anchor points (shown as black dots) that are arranged along a section of a circle. Notice that the hyperplanes tangent to the anchor points achieve all possible bag labelings. The hypothesis h shown above, for example, labels the red and blue bags positive, and the green bag negative.

It seems that to bound $\text{err}(\bar{h})$ we must ensure that a non-negligible portion of a positive bag be labeled positive. A natural way of accomplishing this is to use a real-valued version of the instance hypothesis class (i.e., classifiers of the form $h_r : \mathcal{I} \to [0,1]$, and labels determined by thresholding), and requiring that functions in this class (a) be *smooth*, and (b) label a positive bag with a certain *margin*. To understand why these properties are needed, consider three ways that h_r can label the instances of a positive bag b as one varies the latent parameter α (i.e., x-axis corresponds to instances, y-axis corresponds to classifier output):



In both the left and center panels, h_r labels only a tiny portion of the bag positive: in the first case h_r barely labels any instance above the threshold of $1/2$, resulting in a small margin; in the second case, although the margin is large, h_r changes rapidly along the bag. Finally, in the right panel, when both the margin and smoothness conditions are met, a non-negligible portion of b is labeled positive.

We shall thus study how to bound the generalization error in this setting.

2.2.2. Learning with a Margin

Let $\mathcal{H}_r$ be the real-valued relaxation of $\mathcal{H}$ (i.e. each $h_r \in \mathcal{H}_r$ is now of the form $h_r : \mathcal{I} \to [0,1]$). In order to ensure smoothness we impose a λ-Lipschitz constraint on the instance hypotheses: $\forall h_r \in \mathcal{H}_r$, $x, x' \in \mathcal{I}$, $|h_r(x) - h_r(x')| \leq \lambda \|x - x'\|_2$. We denote the corresponding bag hypothesis class as $\overline{\mathcal{H}}_r$. Note that the true bag labels are still binary in this setting (i.e. determined by h^*).

Similar to the VC-dimension, the "fat-shattering dimension" of a real-valued bag hypothesis class, $\text{FAT}_\gamma(\overline{\mathcal{H}}_r)$, relates the generalization error to the empirical error at margin γ (see for example Anthony & Bartlett, 1999):

$$\widehat{\text{err}}_\gamma(\bar{h}_r, S_m) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^{m} \mathbf{1}\{\text{MARGIN}(\bar{h}_r(b_i), y_i) < \gamma\}, \quad (1)$$

$$\text{where } \text{MARGIN}(x,y) \stackrel{\text{def}}{=} \begin{cases} x - 1/2 & y = 1 \\ 1/2 - x & \text{otherwise} \end{cases}.$$

Recall that it was not possible to bound generalization error in terms of the instance hypotheses using VC dimension. However, analogous to Sabato & Tishby's analysis of finite size bags (2009), we *can* bound generalization error for manifold bags in terms of the fat-shattering dimension of instance hypotheses, $\text{FAT}_\gamma(\mathcal{H})$. In particular, we have the following:

Theorem 3 *Let $\mathcal{B}$ belong to class (V, n, τ). Let $\mathcal{H}_r$ be λ-Lipschitz smooth (w.r.t. ℓ_2-norm), and $\overline{\mathcal{H}}_r$ be the corresponding bag hypotheses over $\mathcal{B}$. Pick any $0 < \gamma < 1$ and $m \geq \text{FAT}_{\gamma/16}(\mathcal{H}_r) \geq 1$. For any $0 < \delta < 1$, we have with probability at least $1 - \delta$ over an i.i.d. sample S_m (of size m), for every $\bar{h}_r \in \overline{\mathcal{H}}_r$:*

$$err(\bar{h}_r) \leq \widehat{err}_\gamma(\bar{h}_r, S_m) + O\left(\sqrt{\frac{n^2 \text{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m} \log^2\left(\frac{Vm}{\gamma^2 \tau_0^n}\right) + \frac{1}{m}\ln\frac{1}{\delta}}\right), \quad (2)$$

where $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$.

Observe that the complexity term in Eq. (2) is independent of the "bag size"; it has instead been replaced by the volume and other geometric properties of the manifold bags. The other term captures the sample error for individual hypotheses at margin γ. Thus a natural strategy for a learner is to return a hypothesis that minimizes the empirical error while maximizing the margin.

2.3. Learning from Queried Instances

So far we have analyzed the MIL learner as a black box entity, which can minimize the empirical bag error by somehow accessing the bags. Since the individual bags in our case are low-dimensional manifolds (with an infinite number of instances), we must also consider *how* these bags are accessed by the learner. Perhaps the simplest approach is to query ρ instances

uniformly from each bag, thereby "reducing" the problem to standard MIL (with finite size bags) for which there are algorithms readily available (e.g. Maron & Lozano-Perez, 1998; Andrews et al., 2002; Zhang & Goldman, 2002; Viola et al., 2005). More formally, for a bag sample S_m, let $p_1^i, \ldots, p_\rho^i$ be ρ independent instance samples drawn uniformly from the (image of) bag $b_i \in S_m$, and let $S_{m,\rho} \stackrel{\text{def}}{=} \bigcup_{i,j} p_j^i$ be the set of all instances. Assuming that our manifold bags have well-conditioned boundaries, the following theorem relates the empirical error of sampled bags, $\widehat{\text{err}}_\gamma(\bar{h}_r, S_{m,\rho}) \stackrel{\text{def}}{=} \frac{1}{m}\sum_{i=1}^m \mathbf{1}\{\text{MARGIN}(\max_{j\in[\rho]} h(p_j^i), y_i) < \gamma\}$, to the generalization error.

Theorem 4 *Let $\mathcal{B}$ belong to class (V, n, τ). Let $\mathcal{H}_r$ be λ-Lipschitz smooth (w.r.t. ℓ_2-norm), and $\overline{\mathcal{H}}_r$ be the corresponding bag hypotheses over $\mathcal{B}$. Pick any $0 < \delta_1, \delta_2 < 1$, then with probability at least $1-\delta_1-\delta_2$, over the draw of m bags (S_m) and ρ instances per bag ($S_{m,\rho}$), for all $\bar{h}_r \in \overline{\mathcal{H}}_r$ we have the following:*

Let $\frac{1}{\kappa} \stackrel{def}{=} \max_{b_i \in S_m}\{\text{COND}(\partial b_i)\}$ (where ∂b_i is the boundary of the manifold bag b_i) and set $\tau_1 = \min\{\frac{\tau}{32}, \frac{\kappa}{8}, \frac{\gamma}{9\lambda}, \frac{\gamma}{9}\}$. If

$$\rho \geq \Omega\left(\left(V/\tau_1^{c_0 n}\right)\left(n + \ln\left(\frac{mV}{\tau_1^n \delta_2}\right)\right)\right),$$

then

$$err(\bar{h}_r) \leq \widehat{err}_{2\gamma}(\bar{h}_r, S_{m,\rho}) + O\left(\sqrt{\frac{n^2 \text{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m}\log^2\left(\frac{Vm}{\gamma^2\tau_0^n}\right) + \frac{1}{m}\ln\frac{1}{\delta_1}}\right),$$

where $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$ and c_0 is an absolute constant.

Notice the effect of the two key parameters in the above theorem: the number of training bags, m, and the number of queried instances per bag, ρ. Increasing either quantity improves generalization – increasing m drives down the error (via the complexity term), while increasing ρ helps improve the confidence (via δ_2). While ideally we would like both quantities to be large, increasing these parameters is, of course, computationally burdensome for a standard MIL learner. Note, however, the difference between m and ρ: increasing m comes at an *additional cost* of obtaining extra labels, whereas increasing ρ does not. We would therefore like an algorithm that can take advantage of using a large ρ while avoiding computational costs.

2.3.1. Iterative Querying Heuristic

As we saw in the previous section, we would ideally like to train with a large number of queried instances, ρ, per training bag. However, this may be impractical in terms of both speed and memory constraints. Suppose we have access to a black box MIL algorithm $\mathcal{A}$ that can only train with $\hat{\rho} < \rho$ instances per bag at once. We propose a procedure called Iterative Querying Heuristic (IQH), described in detail in Algorithm 1 (the main steps are highlighted in blue).

Algorithm 1 Iterative Querying Heuristic (IQH)

Input: Training bags $(b_1, \ldots, b_m)$, labels $(y_1, \ldots, y_m)$, parameters T, ω and $\hat{\rho}$

1: Initialize $I_i^0 = \emptyset$, h_r^0 as any classifier in $\mathcal{H}_r$.
2: **for** $t = 1, \ldots, T$ **do**
3: **Query ω new candidate instances per bag**: $Z_i^t := I_i^{t-1} \cup \{p_1^i, \ldots, p_\omega^i\}$ where $p_j^i \sim b_i$, $\forall i$.
4: **Keep $\hat{\rho}$ highest scoring inst. using h_r^{t-1}**: $I_i^t \subset Z_i^t$ s.t. $|I_i^t| = \hat{\rho}$ and $h_r^{t-1}(p) \geq h_r^{t-1}(p')$ for all $p \in I_i^t, p' \in Z_i^t \setminus I_i^t$.
5: **Train $\bar{h}_r^t$ with the selected instances**: $\bar{h}_r^t \leftarrow \mathcal{A}(\{I_1^t \ldots I_m^t\}, \{y_1 \ldots y_m\})$.
6: **end for**
7: **Return** h_r^T and the corresponding $\bar{h}_r^T$

Notice that IQH uses a total of $T\hat{\rho}$ instances per bag for training (T iterations times $\hat{\rho}$ instances per iteration). Thus, setting $T \approx \rho/\hat{\rho}$ should achieve performance comparable to using ρ instances at once. The free parameter ω controls how many new instances are considered in each iteration.

The intuition behind IQH is as follows. For positive bags, we want to ensure that at least one of the queried instances is positive; hence we use the current estimate of the classifier to select the most positive instances. For negative bags, we know all instances are negative. In this case we select the instances that are closest to the decision boundary of our current classifier (corresponding to the most difficult negative instances); the motivation for this is similar to bootstrapping negative examples (Felzenszwalb et al., 2009) and some active learning techniques (Cohn et al., 1994). We then use these selected instances to find a better classifier.

Thus one expects IQH to take advantage of a large number of instances per bag, without actually having to train with all of them at one time.

3. Experiments

Recall that we have shown that the generalization error is bounded in terms of key geometric properties of the manifold bags, such as curvature ($1/\tau$) and volume (V). Here we will experimentally validate that generalization error does indeed scale with these quantities, providing an empirical lower bound. Additionally, we

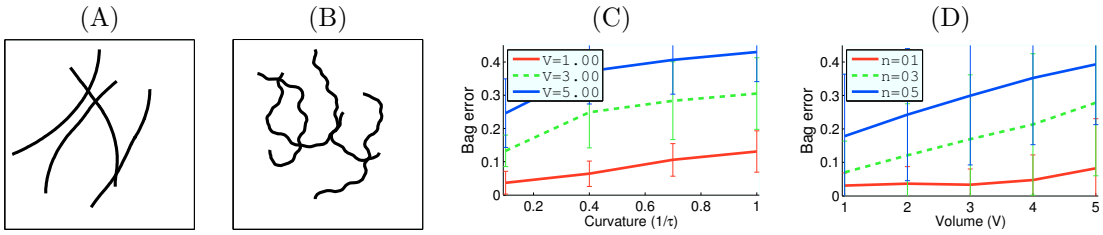


Figure 3. **Synthetic Data Results:** Examples of four synthetically generated bags in $\mathbb{R}^2$ with **(A)** low curvature and **(B)** high curvature. **(C)** and **(D)**: Test error scales with the manifold parameters: volume (V), curvature ($\frac{1}{\tau}$), and dimension (n).

study how the choice of ρ affects the error, and show that our Iterative Heuristic (IQH) is effective in reducing the number of instances needed to train in each iteration. In all our experiments we use a boosting algorithm for MIL called MILBoost (Viola et al., 2005) as the black box $\mathcal{A}$; additional experiments with the MI-SVM algorithm (Andrews et al., 2002) are available in Appendix C. Both algorithms show similar trends and we expect the same for any other choice of $\mathcal{A}$. Note that we use IQH only where specified.

3.1. Synthetic Data

We begin with a carefully designed synthetic dataset, where we have complete control over the manifold curvature, volume and dimension, and study its effects on the generalization. The details on how we generate the dataset are provided in Appendix B; see Figure 3 (A) and (B) for examples of the generated manifolds.

For the first set of experiments, we study the interplay between the volume and curvatue while keeping the manifold dimension fixed. Here we generated one-dimensional curves of specified volume (V) and curvature ($1/\tau$) in $\mathbb{R}^2$. We set h^* to be a vertical hyperplane and labeled the samples accordingly (see Appendix B). For training, we used 10 positive and 10 negative bags with 500 queried instances per bag (forming a good cover); for testing we used 100 bags. Figure 3 (C) shows the test error, averaged over 50 trials, as we vary these parameters. Observe that for a fixed V, as we increase $1/\tau$ (making the manifolds more curvy) generalization error goes up.

For the next set of experiments, we want to understand how manifold dimensionality affects the error. Here we set the ambient dimension to 10 and varied the manifold dimension (with all other experiment settings as before). Figure 3 (D) shows how the test error scales for different dimensional bags as we vary the volume ($1/\tau$ set to 1).

These results corroborate the general intuition of our analysis, and give an empirical verification that the error indeed scales with the geometric properties of a manifold bag.

3.2. Real Data

In this section we present results on image and audio datasets. We will see that the generalization behavior is consistent with our analysis across these different domains. We also study the effects of varying ρ on generalization error, and see how using IQH helps achieve similar error rates with less instances per iteration.

Figure 4. **INRIA Heads:** for our experiments we have labeled the heads in the INRIA Pedestrian Dataset (Dalal & Triggs, 2005). We can construct bags of different volume by padding the head region. The above figure shows positive bags for two different amounts of padding.

INRIA Heads. For these experiments we chose the task of head detection (e.g. positive bags are images which contain at least one head). We used the INRIA Pedestrian Dataset (Dalal & Triggs, 2005), which contains both pedestrian and non-pedestrian images, to create an INRIA Heads dataset as follows. We manually labeled the location of the head in the pedestrian images. The images were resized such that the size of the head is roughly 24×24 pixels; therefore, instances in this experiment are image patches of that size. For each image patch we computed Haar-like features on various channels as in (Dollár et al., 2009), which corresponds to our instance space $\mathcal{I}$.

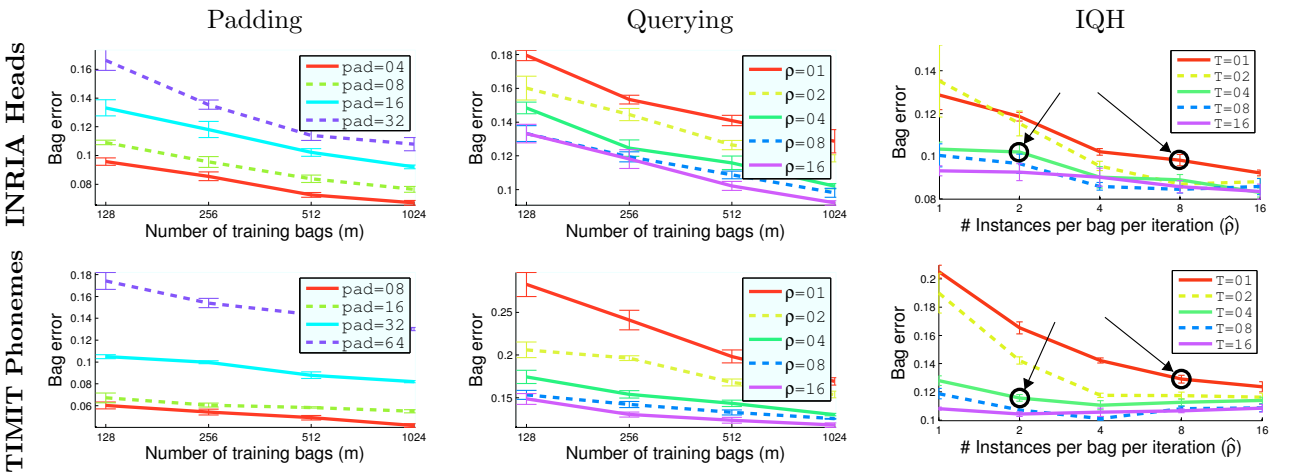


Figure 5. **Image and Audio Results:** three different experiments (columns) – varying padding (volume), number of queried instances, and number of IQH iterations – on two different datasets (rows); see text for details. Note that x-axes are in logarithmic scale. All reported results are averages over 5 trials.

Using the ground truth labels, we generated 2472 positive bags by cropping out the head region with different amounts of padding (see Figure 4), which corresponds to changing the volume of the manifold bags. For example, padding by 6 pixels results in a bag that is a 30×30 pixel image. To generate negative bags we cropped 2000 random patches from the non-pedestrian images, as well as non-head regions from the pedestrian images. Unless otherwise specified, padding was set to 16.

TIMIT Phonemes. Our other application is in the audio domain, and is analogous to the image data described above. The task here was to detect whether a particular phoneme is spoken in an audio clip (we arbitrarily chose the phoneme "s" to be the positive class). We used the TIMIT dataset (Garofolo et al., 1993), which contains recordings of over 600 speakers reading text; the dataset also contains phoneme annotations. Bags in this experiment are audio clips, and instances are audio pieces of length 0.2 seconds (i.e. this is the size of our sliding window). As in the image experiments, we had ground truth annotation for instances, and generated bags of various volumes/lengths by padding. We computed features as follows: we split each sliding window into 25 millisecond pieces, computed Mel-frequency cepstral coefficients (MFCC) (Davis & Mermelstein, 1980; Ellis, 2005) for each piece, and concatenated them to form a 104 dimensional feature vector for each instance. The reported padding amounts are in terms of a 5 millisecond step size (e.g., padding of 8 corresponds to 40 milliseconds of concatenation). Unless otherwise specified, padding was set to 64.

Results. Our first set of experiments involved sweeping over the amount of padding (corresponding to varying the volume of bags). We train with a fixed number of instances per bag, $\rho = 4$. Results for different training set sizes (m) are shown in the first column of Figure 5. As observed in the synthetic experiments, we see that increasing the padding (volume) leads to poorer generalization for both datasets. This corroborates our basic intuition that learning becomes more difficult with manifolds of larger volume.

In our second set of experiments, the goal was to see how generalization error is affected by varying the number of queried instances per bag, which compliments Theorem 4. Results are shown in the middle column of Figure 5. Observe the interplay between m and ρ: increasing either, while keeping the other fixed, drives the error down. Recall, however, that increasing m also requires additional labels while querying more instances per bag does not. The number of instances indeed has a significant impact on generalization – for example, in the audio domain, querying more instances per bag can improve the error by up to 15%. As per our analysis, these results suggest that to fully leverage the training data, we must query many instances per bag. Since training with a large number of instances can become computationally prohibitive, this further justifies the Iterative Querying Heuristic (IQH) described in Section 2.3.1.

Our final set of experiments evaluates the proposed IQH method (see Algorithm 1). The number of training bags, m, was fixed to 1024, and the number of candidate instances per iteration, ω, was fixed to 32 for both datasets. Note that $T = 1$ corresponds to

querying instances and training MILBoost once (i.e. no iterative querying). Results are shown in the right column of Figure 5. These results show that our heuristic works quite well. Consider the highlighed points in both plots: using IQH with $T = 4$ and just 2 instances per bag during training we are able to achieve comparable test error to the naive method (i.e. $T = 1$) with 8 instances per bag. Thus, using IQH, we can obtain a good classifier while needing to use less memory and computational resources per iteration.

4. Conclusion

We have presented a new formulation of MIL where bags are manifolds in the instance space, rather than finite sets of instances. This scenario often appears in practice, but has thus far been overlooked in theoretical analysis and algorithm design. We showed that manifold geometry is intimately related to PAC-learnability for this formulation. Our experimental results corroborate the basic intuition of our analysis. Our iterative querying technique enables us to achieve good generalization error while needing to use less memory and computational resources, and should thus be of immediate practical value. We hope that our work encourages further research into leveraging manifold structure in designing MIL algorithms.

Acknowledgments

B.B. was supported by a 2010 Google Fellowship. N.V. was supported in part by NSF Grant CNS-0932403. S.B. was supported by ONR MURI Grant #N00014-08-1-0638 and NSF Grant AGS-0941760.

References

Ali, S. and Shah, M. Human action recognition in videos using kinematic features and multiple instance learning. *PAMI*, 2008.

Andrews, S., Hofmann, T., and Tsochantaridis, I. Multiple instance learning with generalized support vector machines. *A.I.*, pp. 943–944, 2002.

Anthony, M. and Bartlett, P.L. *Neural network learning: Theoretical foundations*. Cambridge Univ Pr., 1999.

Auer, P., Long, P. M., and Srinivasan, A. Approximating hyper-rectangles: Learning and pseudo-random sets. In *Proc. of ACM Symposium on Theory of Comp.*, 1997.

Blum, A. and Kalai, A. A Note on Learning from Multiple-Instance Examples. *Mach. Learning*, 30(1):23–29, 1998.

Buehler, P., Zisserman, A., and Everingham, M. Learning sign language by watching TV (using weakly aligned subtitles). In *CVPR*, 2009.

Clarkson, K. Tighter bounds for random projections of manifolds. *Comp. Geometry*, 2007.

Cohn, D., Atlas, L., and Ladner, R. Improving generalization with active learning. *Machine Learning*, 1994.

Dalal, N. and Triggs, B. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

Davis, S. and Mermelstein, P. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE TASSP*, 1980.

Dietterich, T.G., Lathrop, R.H., and Lozano-Perez, T. Solving the multiple-instance problem with axis parallel rectangles. *A.I.*, 1997.

Dollár, P., Tu, Z., Perona, P., and Belongie, S. Integral channel features. In *BMVC*, 2009.

Ellis, D.P.W. PLP and RASTA (and MFCC, and inversion) in Matlab. `http://www.ee.columbia.edu/~dpwe/resources/matlab/rastamat/`, 2005.

Felzenszwalb, P., Girshick, R., McAllester, D., and Ramanan, D. Object Detection with Discriminatively Trained Part-Based Models. *PAMI*, 2009.

Garofolo, J.S. et al. TIMIT Acoustic-Phonetic Continuous Speech Corpus. `http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC93S1`, 1993.

Long, P.M. and Tan, L. PAC Learning Axis-aligned Rectangles with Respect to Product Distributions from Multiple-Instance Examples. *Machine Learning*, 1998.

Mandel, M.I. and Ellis, D.P.W. Multiple-instance learning for music information retrieval. In *ISMIR*, 2008.

Maron, O. and Lozano-Perez, T. A framework for multiple-instance learning. In *NIPS*, 1998.

Niyogi, P., Smale, S., and Weinberger, S. Finding the homology of submanifolds with high confidence from random samples. *Disc. Computational Geometry*, 2006.

Sabato, S. and Tishby, N. Homogeneous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

Sabato, S., Srebro, N., and Tishby, N. Reducing Label Complexity by Learning From Bags. In *AISTATS*, 2010.

Saul, L.K., Rahim, M.G., and Allen, J.B. A statistical model for robust integration of narrowband cues in speech. *Comp. Speech and Language*, 15:175–194, 2001.

Stikic, M. and Schiele, B. Activity recognition from sparsely labeled data using multi-instance learning. In *Location and Context Awareness*. 2009.

Vapnik, V.N. and Chervonenkis, A.Y. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Prob. and Its App.*, 1971.

Viola, P., Platt, J.C., and Zhang, C. Multiple instance boosting for object detection. In *NIPS*, 2005.

Zhang, Q. and Goldman, S.A. EM-DD: An improved multiple-instance learning technique. In *NIPS*, 2002.

A. Appendix: Proofs

A.1. Proof of Theorem 2

We will show this for $n = 1$ and $N = 2$ (the generalization to $N > n \geq 1$ is immediate). We first construct 2^m anchor points $p_0, \ldots, p_{2^m-1}$ on a section of a circle in $\mathbb{R}^2$ that will serve as a guide on how to place m bags $b_1, \ldots, b_m$ of dimension $n = 1$, volume[4] $\leq V$, and condition number $\leq 1/\tau$ in $\mathbb{R}^2$. We will then show that the class of hyperplanes in $\mathbb{R}^2$ can realize all possible 2^m labelings of these m bags.

Let $V_0 \stackrel{\text{def}}{=} \min(V/2, \pi)$. Define anchor points $p_i \stackrel{\text{def}}{=} (2\tau\cos(\frac{V_0 i}{2\tau 2^m}), 2\tau\sin(\frac{V_0 i}{2\tau 2^m}))$ for $0 \leq i \leq 2^m - 1$. Observe that the points p_i are on a circle centered at the origin of radius 2τ in $\mathbb{R}^2$.

We use points $p_0, \ldots, p_{2^m-1}$ as guides to place m bags $b_1, \ldots, b_m$ in $\mathbb{R}^2$ that are contained entirely in the disc of radius 2τ centered at the origin and pass through the anchor points as follows. Let $k_m^i \ldots k_1^i$ represent the binary representation of the number i ($0 \leq i \leq 2^m - 1$). Place bag b_j such that b_j passes through the anchor point p_i, if and only if $k_j^i = 1$. (see figure below for a visual example for 3 bags and 8 anchor points). Note that since, by construction, the arc (the dotted line in the figure) containing the anchor points has condition number at most $1/2\tau$ with volume strictly less than V, bags b_j can be made to have condition number at most $1/\tau$ with volume at most V.

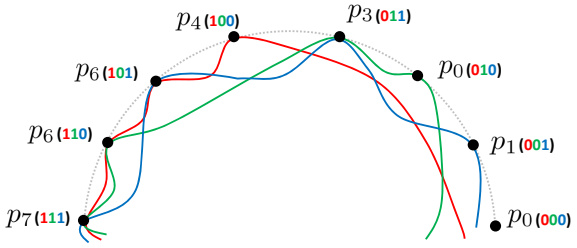


Figure 6. **Placement of arbitrarily smooth bags along a section of a disk**. Three bags (colored blue, green and red) go around the eight anchor points $p_0, \ldots, p_7$ in such a way that the hypothesis class of hyperplanes can realize all possible bag labelings.

It is clear that hyperplanes in $\mathbb{R}^2$ can realize any possible labeling of these m bags. Say, we want some arbitrary labeling $(+1, +1, 0, \ldots, +1)$. We look at the number i with the same bit representation. Then a hyperplane that is tangent to the circle (centered at the origin and radius 2τ) at the anchor point p_i, labels p_i positive, and all other p_k's negative. Note that this hypothesis will also label exactly those bags b_j positive that are passing through the point p_i, and rest of the bags labeled negative. Thus, realizing the arbitrary labeling.

[4]Volume of a 1-dimensional manifold is its length.

A.2. Proof of Theorem 3

Before stating the proof, we give the following useful fact about manifolds with bounded volume and curvature.

Fact 5 [manifold covers – see Section 2.4 of (Clarkson, 2007)] *Let $M \subset \mathbb{R}^N$ be a compact n-dimensional manifold with $\text{VOL}(M) \leq V$ and $\text{COND}(M) \leq 1/\tau$. Pick any $0 < \epsilon \leq \tau/2$. There exists an ϵ-covering of M of size at most $2^{c_0 n}(V/\epsilon^n)$, where c_0 is an absolute constant. That is, there exists $C \subset M$ such that $|C| \leq 2^{c_0 n}(V/\epsilon^n)$ with the property: for all $p \in M$, $\exists\ q \in C$ such that $\|p - q\| \leq \epsilon$.*

Now, for any domain X, real-valued hypothesis class $H \subset [0,1]^X$, margin $\gamma > 0$ and a sample $S \subset X$, define

$$\text{cov}_\gamma(H, S) \stackrel{\text{def}}{=} \{C \subset H \mid \forall h \in H, \exists h' \in C, \max_{s \in S} |h(s) - h'(s)| \leq \gamma\}$$

as a set of γ-covers of S by H. Let γ-covering number of H for any integer $m > 0$ be defined as

$$\mathcal{N}_\infty(\gamma, H, m) \stackrel{\text{def}}{=} \max_{S \subset X : |S| = m} \min_{C \in \text{cov}_\gamma(H,S)} |C|.$$

We will first relate the covering numbers of $\mathcal{H}_r$ and $\overline{\mathcal{H}}_r$ with the fat-shattering dimension in the following two lemmas.

Lemma 6 [relating hypothesis cover to the fat-shattering dimension – see Theorem 12.8 (Anthony & Bartlett, 1999)] *Let H be a set of real functions from a domain X to the interval $[0,1]$. Let $\gamma > 0$. Then for $m \geq \text{FAT}_{\gamma/4}(H)$,*

$$\mathcal{N}_\infty(\gamma, H, m) < 2\big(4m/\gamma^2\big)^{\text{FAT}_{\gamma/4}(H)\log\frac{4em}{\text{FAT}_{\gamma/4}(H)\gamma}}.$$

Lemma 7 [adapted from Lemma 17 of (Sabato & Tishby, 2009)] *Let $\mathcal{H}_r$ be an instance hypothesis class such that each $h_r \in \mathcal{H}_r$ is λ-lipschitz (w.r.t. ℓ_2-norm), and let $\overline{\mathcal{H}}_r$ be the corresponding bag hypothesis class over $\mathcal{B}$ that belongs to the class (V, n, τ). For any $\gamma > 0$ and $m \geq 1$, we have*

$$\mathcal{N}_\infty(2\gamma, \overline{\mathcal{H}}_r, m) \leq \mathcal{N}_\infty(\gamma, \mathcal{H}_r, m2^{c_0 n}(V/\epsilon^n)),$$

where $\epsilon = \min\{\frac{\tau}{2}, \frac{\gamma}{2}, \frac{\gamma}{2\lambda}\}$, and c_0 is an absolute constant.

Proof. Let $S = \{b_1, \ldots, b_m\}$ be a set of m manifold bags. Set $\epsilon = \min\{\frac{\tau}{2}, \frac{\gamma}{2}, \frac{\gamma}{2\lambda}\}$. For each bag $b_i \in S$, let C_i be the smallest ϵ-cover of (the image of) b_i (by Fact 5, we know that $|C_i| \leq 2^{\mathrm{c}_0 n}(V/\epsilon^n)$ for some absolute constant c_0).

Define $S^{\cup} \stackrel{\text{def}}{=} \cup_i C_i$ and let $R \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^{\cup})$ be some γ-cover of $S^{\cup}$. Now, for any $h_r \in \mathcal{H}_r$, let $\bar{h}_r \in \overline{\mathcal{H}}_r$ denote the corresponding bag classifier, and define $\widetilde{h}_r(C_i) \stackrel{\text{def}}{=} \max_{c \in C_i} h_r(c)$ as the maximum attained by h_r on the sample C_i. Then, since h_r is λ-lipschitz (w.r.t. ℓ_2-norm), we have for any bag b_i and its corresponding ϵ-cover C_i,

$$|\bar{h}_r(b_i) - \widetilde{h}_r(C_i)| \leq \lambda\epsilon.$$

It follows that $\forall x \in S^{\cup}$: for any $h_r \in \mathcal{H}_r$ and $h'_r \in R$ such that $|h_r(x) - h'_r(x)| \leq \gamma$ (and the corresponding bag classifiers $\bar{h}_r$ and $\bar{h}'_r$ in $\overline{\mathcal{H}}_r$),

$$\begin{aligned}
&\max_{i\in[m]} |\bar{h}_r(b_i) - \bar{h}'_r(b_i)| \\
&\quad= \max_{i\in[m]} |\bar{h}_r(b_i) - \widetilde{h}_r(C_i) + \widetilde{h}_r(C_i) - \widetilde{h}'_r(C_i) \\
&\qquad\qquad + \widetilde{h}'_r(C_i) - \bar{h}'_r(b_i)| \\
&\quad\leq 2\lambda\epsilon + \gamma \leq 2\gamma.
\end{aligned}$$

Also, note that for any $h_r \in \mathcal{H}_r$ and $h'_r \in R$ such that $|h_r(x) - h'_r(x)| \leq \gamma$ ($x \in S^{\cup}$), we have $\bar{h}'_r \in \{\bar{h}_r | h_r \in R\} \stackrel{\text{def}}{=} \bar{R}$. It follows that for any $R \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^{\cup})$, $\{\bar{h}_r | h_r \in R\} \in \mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r, S)$. Thus,

$$\{\overline{H}_r | H_r \in \mathrm{cov}_\gamma(\mathcal{H}_r, S^{\cup})\} \subset \mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r, S).$$

Hence, we have

$$\begin{aligned}
\mathcal{N}_\infty(2\gamma, \overline{\mathcal{H}}_r, m) &= \max_{S\subset\mathcal{B}:|S|=m} \min_{\bar{R}\in\mathrm{cov}_{2\gamma}(\overline{\mathcal{H}}_r,S)} |\bar{R}| \\
&\leq \max_{S\subset\mathcal{B}:|S|=m} \min_{\bar{R}\in\{\bar{H}_r|H_r\in\mathrm{cov}_\gamma(\mathcal{H}_r,S^{\cup})\}} |\bar{R}| \\
&= \max_{S\subset\mathcal{B}:|S|=m} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S^{\cup})} |R| \\
&= \max_{S\subset\mathcal{I}:|S|=|S^{\cup}|m} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S)} |R| \\
&\leq \max_{S\subset\mathcal{I}:|S|=m2^{\mathrm{c}_0 n}(V/\epsilon^n)} \min_{R\in\mathrm{cov}_\gamma(\mathcal{H}_r,S)} |R| \\
&= \mathcal{N}_\infty(\gamma, \mathcal{H}_r, m2^{\mathrm{c}_0 n}(V/\epsilon^n)),
\end{aligned}$$

where c_0 is an absolute constant. ∎

Now we can relate the empirical error with generalization error by noting the following lemma.

Lemma 8 [generalization error bound for real-valued functions – Theorem 10.1 of (Anthony & Bartlett, 1999)] *Suppose that F is a set of real-valued functions defined on the domain X. Let $\mathcal{D}$ be any probability distribution on $Z = X \times \{0, 1\}$, $0 \leq \epsilon \leq 1$, real $\gamma > 0$ and integer $m \geq 1$. Then,*

$$\begin{aligned}
&\Pr_{S_m\sim\mathcal{D}}\left[\exists f \in F : err(f) \geq \widehat{err}_\gamma(f, S_m) + \epsilon\right] \\
&\qquad\qquad \leq 2\mathcal{N}_\infty\Big(\frac{\gamma}{2}, F, 2m\Big)e^{-\epsilon^2 m/8},
\end{aligned}$$

where S_m is an i.i.d. sample of size m from $\mathcal{D}$, $err(f)$ is the error of f with respect to $\mathcal{D}$, and $\widehat{err}_\gamma(f, S_m)$ is the empirical error of f with respect to S_m at margin γ.

Combining Lemmas 8, 7 and 6, we have (for $m \geq \mathrm{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)$):

$$\begin{aligned}
&\Pr_{S_m\sim\mathcal{D}_\mathcal{B}}\left[\exists \bar{h}_r \in \overline{\mathcal{H}}_r : \mathrm{err}(\bar{h}_r) \geq \widehat{\mathrm{err}}_\gamma(\bar{h}_r, S_m) + \epsilon\right] \\
&\quad\leq\ 2\,\mathcal{N}_\infty\Big(\frac{\gamma}{2}, \overline{\mathcal{H}}_r, 2m\Big)\, e^{-\epsilon^2 m/8} \\
&\quad\leq\ 2\,\mathcal{N}_\infty\Big(\frac{\gamma}{4}, \mathcal{H}_r, m2^{\mathrm{c}_0 n}(V/\tau_0^n)\Big)\, e^{-\epsilon^2 m/8} \\
&\quad\leq\ 4\Big(\frac{64\cdot 2^{\mathrm{c}_0 n} Vm}{\gamma^2\tau_0^n}\Big)^{d\log\left(\frac{16\epsilon 2^{\mathrm{c}_0 n} Vm}{\tau_0^n d\gamma}\right)}\, e^{-\epsilon^2 m/8},
\end{aligned}$$

where c_0 is an absolute constant, $d \stackrel{\text{def}}{=} \mathrm{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)$, and $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$. For $d \geq 1$, the theorem follows.

A.3. Proof of Theorem 4

We start with the following useful observations that will help in our proof.

Notation: for any two points p and q on a Riemannian manifold M,

- let $D_G(p, q)$ denote the geodesic distance between points p and q.
- $B_G(p, \epsilon) \stackrel{\text{def}}{=} \{p' \in M | D_G(p', p) \leq \epsilon\}$ denote the geodesic ball centered at p of radius ϵ.

Fact 9 [manifold volumes – see Lemma 5.3 (Niyogi et al., 2006)] *Let $M \subset \mathbb{R}^N$ be a compact n-dimensional manifold with $\mathrm{COND}(M) \leq 1/\tau$. Pick any $p \in M$ and let $A_\epsilon \stackrel{def}{=} M \cap B(p, \epsilon)$, where $B(p, \epsilon)$ is a Euclidean ball in $\mathbb{R}^N$ centered at p of radius ϵ. If A_ϵ does not contain any boundary points of M, then $\mathrm{VOL}(A_\epsilon) \geq (\cos(\arcsin(\epsilon/2\tau)))^n \mathrm{VOL}(B_\epsilon^n)$, where B_ϵ^n is a Euclidean ball in $\mathbb{R}^n$ of radius ϵ. In particular, noting that $\mathrm{VOL}(B_\epsilon^n) \geq \epsilon^{\mathrm{c}_0 n}$ for some absolute constant c_0, if $\epsilon \leq \tau$, we have $\mathrm{VOL}(A_\epsilon) \geq \epsilon^{\mathrm{c}_0 n}$.*

Fact 10 [relating geodesic distances to ambient Euclidean distances – see Prop. 6.3 (Niyogi et al., 2006)] *Let $M \subset \mathbb{R}^N$ be a compact manifold with $\text{COND}(M) \leq 1/\tau$. If $p, q \in M$ such that $\|p-q\| \leq \frac{\tau}{2}$, then $D_G(p,q) \leq 2\|p-q\|$.*

Lemma 11 *Let $M \subset \mathbb{R}^N$ be a compact n-dimensional manifold with $\text{VOL}(M) \leq V$ and $\text{COND}(M) \leq 1/\tau$. Let $\mu(M)$ denote the uniform probability measure over M. Define $\mathcal{F}(M,\epsilon) \stackrel{def}{=} \{B_G(p,\epsilon) : p \in M$ and $B_G(p,\epsilon)$ contains no points from the boundary of $M\}$, that is, the set of all geodesic balls of radius ϵ that are contained entirely in the interior of M. Let $\tau_0 \leq \tau$ and $\rho \geq 1$. Let $p_1, \ldots, p_\rho$ be ρ independent draws from $\mu(M)$. Then,*

$$\Pr_{p_1,\ldots,p_\rho \sim \mu(M)} \left[\exists F \in \mathcal{F}(M,\tau_0) : \forall i, p_i \notin F\right] \leq 2^{c_0 n}(V/\tau_0^n) e^{-\rho(\tau_0^{c_0 n}/V)},$$

where c_0 is an absolute constant.

Proof. Let M° denote the interior of M (i.e., it contains all points of M that are not at the boundary). Let $q_0 \in M$ be any fixed point such that $B_G(q_0, \frac{\tau_0}{2}) \subset M^\circ$. Then, by Facts 9 and 10 we know that $\text{VOL}(B_G(q_0, \frac{\tau_0}{2})) \geq \tau_0^{c_0 n}$. Observing that M has volume at most V, we immediately get that $B_G(q_0, \frac{\tau_0}{2})$ occupies at least $\tau_0^{c_0 n}/V$ fraction of M. Thus

$$\Pr_{p_1,\ldots,p_\rho \sim \mu(M)} \left[\forall i, p_i \notin B_G\big(q_0, \frac{\tau_0}{2}\big)\right] \leq \Big(1 - \frac{\tau_0^{c_0 n}}{V}\Big)^\rho.$$

Now, let $C \subset M$ be a $(\frac{\tau_0}{2})$-geodesic covering of M. Using Facts 5 and 10, we can have $|C| \leq 2^{c_1 n}(V/\tau_0^n)$ (where c_1 is an absolute constant). Define $C' \subset C$ as the set $\{c \in C : B_G(c, \frac{\tau_0}{2}) \subset M^\circ\}$. Then by union bounding over points in C', we have

$$\Pr_{p_1,\ldots,p_\rho \sim \mu(M)} \left[\exists c \in C' : \forall i, p_i \notin B_G\big(c, \frac{\tau_0}{2}\big)\right] \leq |C'|\Big(1 - \frac{\tau_0^{c_0 n}}{V}\Big)^\rho.$$

Equivalently we can say that, with probability at least $1 - |C'|e^{-\tau_0^{c_0 n}\rho/V}$, for all $c' \in C'$, there exists $p_i \in \{p_1, \ldots, p_\rho\}$ such that $p_i \in B_G(c', \frac{\tau_0}{2})$.

Now, pick *any* $F \in \mathcal{F}(M,\tau_0)$, and let $q \in M$ denote its center (i.e., q such that $B_G(q,\tau_0) = F$). Then since C is a $(\frac{\tau_0}{2})$-geodesic cover of M, there exists $c \in C$ such that $D_G(q,c) \leq \tau_0/2$. Also, note that c belongs to the set C', since $B_G(c, \tau_0/2) \subset B_G(q,\tau_0) = F \subset M^\circ$. Thus with probability $\geq 1 - |C'|e^{-\tau_0^{c_0 n}\rho/V}$, there exists p_i such that

$$p_i \in B_G(c, \tau_0/2) \subset B_G(q,\tau_0) = F.$$

Observe that since the choice of F was arbitrary, we have that for any $F \in \mathcal{F}$ (uniformly), there exists $p_i \in \{p_1, \ldots, p_\rho\}$ such that $p_i \in F$. The lemma follows. ∎

Lemma 12 *Let $\mathcal{B}$ belong to class (V, n, τ). Fix a sample of size m $\{b_1, \ldots, b_m\} \stackrel{def}{=} S_m \subset \mathcal{B}$, and let ∂b_i denote the boundary of the manifold bag $b_i \in S_m$. Define $\frac{1}{\kappa} \stackrel{def}{=} \max_{b_i \in S_m}\{\text{COND}(\partial b_i)\}$. Now let $p_1^i, \ldots, p_\rho^i$ be the ρ independent instances drawn uniformly from (the image of) b_i. Let $\mathcal{H}_r$ be a λ-lipschitz (w.r.t. ℓ_2-norm) hypothesis class. Then, for any $\epsilon \leq \min\{\frac{\tau}{32}, \frac{\kappa}{8}\}$,*

$$\Pr\Big[\exists h_r \in \mathcal{H}_r, \exists b_i \in S_m : |\bar{h}_r(b_i) - \max_{j \in [\rho]} h_r(b_i(p_j^i))| > 9\epsilon\lambda\Big] \leq m 2^{c_0 n}(V/\epsilon^n) e^{-\rho\epsilon^{c_0 n}/V},$$

where c_0 is an absolute constant.

Proof. Fix a bag $b_i \in S_m$, and let M denote the manifold b_i. Quickly note that $\text{COND}(M) \leq 1/\tau$.

Define $M_{2\epsilon} \stackrel{\text{def}}{=} \{p \in M : \min_{q \in \partial M} D_G(p,q) \geq 2\epsilon\}$. By recalling that $\text{COND}(\partial M) \leq \frac{1}{\kappa}$ and $\epsilon \leq \min\{\frac{\tau}{32}, \frac{\kappa}{8}\}$, it follows that i) $M_{2\epsilon}$ is non-empty, ii) $\forall x \in M \setminus M_{2\epsilon}$, $\min_{y \in M_{2\epsilon}} D_G(x,y) \leq 8\epsilon$.

Observe that for all $p \in M_{2\epsilon}$, $B_G(p,\epsilon)$ is in the interior of M. Thus by applying Lemma 11, we have:

$$\Pr_{p_1,\ldots,p_\rho \sim \mu(M)} \left[\exists p \in M_{2\epsilon} : \forall i, p_i \notin B_G(p,\epsilon)\right] \leq 2^{c_0 n}(V/\epsilon^n) e^{-\rho(\epsilon^{c_0 n}/V)},$$

where $\mu(M)$ denotes the uniform probability measure on M.

Now for *any* $h_r \in \mathcal{H}_r$, let $x^* \stackrel{\text{def}}{=} \arg\max_{p \in M} h_r(p)$. Then with the same failure probability, we have that there exists some $p_i \in \{p_1, \ldots, p_\rho\}$ such that $D_G(p_i, x^*) \leq 9\epsilon$. To see this, consider:

if $x^* \in M_{2\epsilon}$, $D_G(x^*, p_i) \leq \epsilon$ (for some $p_i \in \{p_1, \ldots, p_\rho\}$), otherwise if $x^* \in M \setminus M_{2\epsilon}$, then exists $q \in M_{2\epsilon}$ such that $D_G(x^*, q) \leq 8\epsilon$.

Noting that h_r is λ-Lipschitz, and union bounding over m bags, the lemma follows. ∎

By Theorem 3 we have for any $0 < \gamma < 1$, with probability at least $1 - \delta_1$ over the sample S_m, for every $\bar{h}_r \in \overline{\mathcal{H}}_r$:

$$\text{err}(\bar{h}_r) \leq \widehat{\text{err}}_\gamma(\bar{h}_r, S_m) + O\left(\sqrt{\frac{n^2 \text{FAT}_{\frac{\gamma}{16}}(\mathcal{H}_r)}{m} \log^2\Big(\frac{Vm}{\gamma^2 \tau_0^n}\Big) + \frac{1}{m}\ln\frac{1}{\delta_1}}\right),$$

where $\tau_0 = \min\{\frac{\tau}{2}, \frac{\gamma}{8}, \frac{\gamma}{8\lambda}\}$.

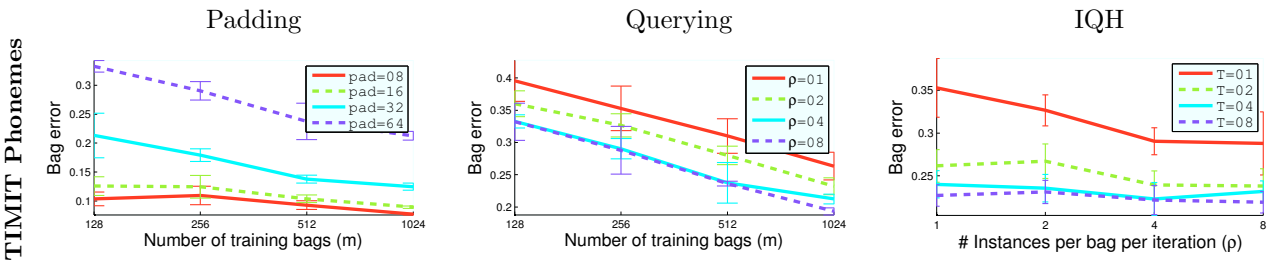


Figure 7. **Audio Results using MI-SVM:** three different experiments (columns) – varying padding (volume), number of queried instances, and number of IQH iterations – on two different datasets (rows); see text for details. Note that x-axes are in logarithmic scale. All reported results are averages over 5 trials.

By applying Lemma 12 (with ϵ set to $\tau_1 = \min\{\frac{\tau}{32}, \frac{\kappa}{8}, \frac{\gamma}{9\lambda}, \frac{\gamma}{9}\}$), it follows that if $\rho \geq \Omega\big((V/\tau_1^{c_0 n})\big(n + \ln\big(\frac{mV}{\tau_1^n \delta_2}\big)\big)\big)$, then with probability $1-\delta_2$: $\widehat{\text{err}}_\gamma(\bar{h}_r, S_m) \leq \widehat{\text{err}}_{2\gamma}(\bar{h}_r, S_{m,\rho})$, yielding the theorem.

B. Appendix: Synthetic Dataset Generation

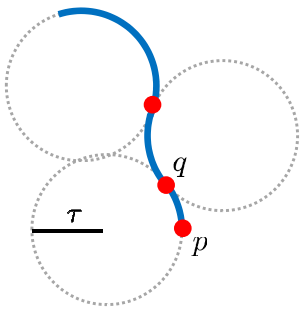


Figure 8. **Synthetic bags.** An example of a synthetic 1-dimensional manifold of a specified volume (length) V and curvature $1/\tau$ generated by our procedure.

We generate a 1-dimensional manifold (in $\mathbb{R}^2$) of curvature $1/\tau$ and volume (length) V as follows (see also Figure 8).

1. Pick a circle with radius τ, a point p on the circle, and a random angle θ (less than π).
2. Choose a direction (either clockwise or counterclockwise) and trace out an arc of length $\theta\tau$ starting at p and ending at point, say, q.
3. Now pick another circle of the same radius τ that is tangent to the original circle at the point q.
4. Repeat the process of tracing out another arc on the new circle, starting at point q and going in the reverse direction.
5. Terminate this process once we have a manifold of volume V.

Notice that this procedure can potentially result in a curve that intersects itself or has the condition number less than $1/\tau$. If this happens, we simply reject such a curve and generate another random curve until we have a well-conditioned manifold.

To generate a higher dimensional manifold, we extend our 1-dimensional manifold ($M \subset \mathbb{R}^2$) in the extra dimensions by taking a Cartesian product with a cube: $M \times [0,1]^{n-1}$. Notice that the "cube"-extension does not alter the condition number (i.e. it remains $1/\tau$). Since the resulting manifold fills up only $n+1$ dimensions, we randomly rotate it the ambient space.

Now, to label the generated manifolds positive and negative, we first fix h^* to be a vertical hyperplane (w.r.t. the first coordinate) in $\mathbb{R}^N$. To label a manifold b negative, we translate it such that the entire manifold lies in the negative region induced by h^*. And to label it positive, we translate it such that a part of b lies in the positive region induced by h^*.

C. Appendix: Additional Experiments with MI-SVM

We repeated the suite of experiments on the TIMIT dataset with the MI-SVM algorithm (Andrews et al., 2002) (implemented using the LIBSVM package[5]). Figure 7 shows the results. Upon comparison with the MILBoost results in Figure 5 we observe that the general trends are quite similar. This reinforces the fact that our results should generalize to most MIL algorithms.

[5] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`