Proceedings of Machine Learning Research vol 65:1–26, 2017

Effective Semi-supervised Learning on Manifolds

Amir Globerson GAMIR@TAU.AC.IL
School of Computer Science, Tel Aviv University.

Roi Livni RLIVNI@CS.PRINCETON.EDU
School of Computer Science, Princeton University.

Shai Shalev-Shwartz SHAIS@CS.HUJI.AC.IL
School of Computer Science and Engineering, Hebrew University.

Abstract

The abundance of unlabeled data makes semi-supervised learning (SSL) an attractive approach for improving the accuracy of learning systems. However, we are still far from a complete theoretical understanding of the benefits of this learning scenario in terms of sample complexity. In particular, for many natural learning settings it can in fact be shown that SSL does not improve sample complexity. Thus far, the only case where SSL provably helps, without compatibility assumptions, is a recent combinatorial construction of Darnstädt et al. (2013). Deriving similar theoretical guarantees for more commonly used approaches to SSL remains a challenge. Here, we provide the first analysis of manifold based SSL, where there is a provable gap between supervised learning and SSL, and this gap can be arbitrarily large. Proving the required lower bound is a technical challenge, involving tools from geometric measure theory. The algorithm we analyze is similar to subspace clustering, and thus our results demonstrate that this method can be used to improve sample complexity.

1. Introduction

Supervised learning is a powerful framework in machine learning, both from a theoretical and practical standpoint. However, it is clear there is much to be gained from using unlabeled data in learning, as the volume of unlabeled data is several orders of magnitude larger than labeled data. The *semi-supervised learning* (SSL) setting (Chapelle et al., 2006) is further motivated by the fact that humans can learn from relatively weak supervision.

From a learning theoretic perspective, supervised learning (SL) is quite well understood, in terms of setting lower and upper bounds on the sample complexity of learning a given hypothesis class (e.g., see Devroye et al., 1996). For semi-supervised learning, the situation is not as clear. Arguably the most interesting question in this context is how an SSL learner can exploit unlabeled data to reduce its *labeled sample complexity*. In this work we describe an SSL learner with provable upper and lower bounds, that demonstrate the clear effect of using unlabeled data. Most previous work mainly focused on obtaining bounds under so called *compatibility* assumptions on the relation between true target function and distribution (e.g., see Balcan and Blum, 2010). Here we prove that SSL can be effective without making such assumptions.

One of the most wide-spread approaches to semi-supervised learning, and our focus in this paper, is the class of *manifold based methods*. In these, the data is assumed to lie on a low dimensional manifold (e.g., Riemannian), and the classifier is constrained to be smooth on this manifold. There

are many instantiations of this idea, but all share the assumption that proximity on a manifold implies similar classification (e.g., see Belkin and Niyogi, 2004; Blum and Chawla, 2001; Lafferty et al., 2004; Rifai et al., 2011; Altun et al., 2006; Yang et al., 2016).[1]

The intuition behind SSL approaches is that knowledge of the feature distribution should reduce the need for labeled data. Presumably, this should result in sample complexity bounds which demonstrate that learning with SSL reduces the sample complexity of supervised learning. However, despite years of research, such results for manifold based learning have still not been obtained. A key technical difficulty in this context is to show that a supervised learner cannot effectively exploit the same structure learned by the SSL learner. While in the strictly supervised setting such lower bounds are fairly standard using the probabilistic method (e.g., see Devroye et al., 1996) they are considerably more complex in the SSL case, as discussed later.

Here we provide the first sample complexity analysis of a *manifold based* SSL learner. Specifically, we show that for a hypothesis class based on algebraic manifolds, SSL has unboundedly better sample complexity than SL. Such results are rare in the literature, and in fact it was hypothesized by Ben-David et al. (2008) that they are inherently unachievable. However, Darnstädt et al. (2013) recently gave a discrete-structured hypothesis class where such a gap also exists. Our example is in a geometric setting and our learner is closer in spirit to methods that perform unsupervised feature learning and attempt to construct a new representation of the data.

Our SSL algorithm is similar to subspace clustering algorithms such as Vidal (2010) and Elhamifar and Vidal (2009). In these, one models a set of points as a union of hyperplanes. Here we show that such an approach can serve as a component of a provably effective SSL scheme.

Our proof techniques involve technical novelties that might be of independent interest. Namely, a combination of the coarea formula and a probabilistic argument. Together, they can be used to derive distribution dependent lower bounds for supervised learning, in a continuous feature space.

Finally, we complement our theoretical findings with a synthetic experiment that demonstrates the sample complexity gap.

2. Related Work

We briefly review relevant sample complexity results in SSL (see also Zhu, 2006, for a survey on theory and practice of SSL). Such results typically require some compatibility assumptions, as discussed below.

Under certain generative assumptions on the joint distribution of features and labels, one can demonstrate positive results. For example Castelli and Cover (1995) analyze the mixture of Gaussians case, and demonstrate a significant effect for observing unlabeled data. This analysis is refined in Singh et al. (2009) and Azizyan et al. (2013) where finite sample complexity bounds are provided. We stress that all these works assume that the true distribution over input features x (denoted by $P(X)$ in what follows), and the true target function are *compatible*. In other words, only certain pairs of distributions and functions are allowed. Our paper deals with a worst case analysis where no such assumptions are made. Namely, we assume that the hypothesis class is fixed and independent of $P(X)$.

Compatibility assumptions are also employed in Niyogi (2008), who constructs a collection of problems where an SSL provably outperforms an SL learner. The construction makes a strong com-

1. Other methods, such as PCA or more recent autoencoders (Kingma and Welling, 2014) also seek a low dimensional representation of the input, but do not explicitly enforce classifier smoothness.

patibility assumption, which significantly reduces the need for labeled data. Another positive result was provided in Ben-David and Urner (2014). They showed that when labeling is deterministic, an algorithm can benefit from unlabeled data if the true target function is not realizable.

So far most of the theoretical analysis of SSL has relied on the assumption that the unlabeled distribution reveals some information about the target function. This approach has been elegantly formulated in Balcan and Blum (2010) who analyze sample complexity bounds of empirical risk minimization schemes that use compatibility information. Intuitively, knowledge about the labeled distribution allows one to reduce the size of the hypothesis space to search over, resulting in better sample complexity.

Perhaps surprisingly, there has also been a flurry of negative results on SSL (Ben-David et al., 2008; Lafferty and Wasserman, 2007; Rigollet, 2006). These show that in many settings SSL has no sample complexity advantage over a supervised learner. Even more surprisingly, Cozman et al. (2003) demonstrate a case where SL can be superior to SSL.

The work that is closest to ours is Darnstädt et al. (2013), which constructs an example where unlabelled data helps in classification, without using compatibility assumptions. Specifically, the authors consider a discrete domain, and show that for any class there exists some distribution where the learner benefits from unlabelled data. While the work of Darnstädt et al. (2013) nicely demonstrates that one can benefit from unlabelled data, the distributions over X for which this result holds do not map to a geometrically intuitive structure. This is precisely where our motivation and technique markedly differ from Darnstädt et al. (2013). Namely, we consider classification functions and distributions that have a clear geometric interpretation, and are closely related to manifold based SSL methods. Specifically, the distributions we consider are inherently continuous in that any specific x will always have measure zero, in contrast to Darnstädt et al. (2013), where this will not be the case.

Our construction and algorithm rely on algebraic manifolds. These have been studied in several machine learning and statistics related works (e.g., Király et al., 2012; Gibilisco, 2010; Drton et al., 2008; Watanabe, 2009). Unsupervised learning of algebraic sets was suggested in Vidal et al. (2005). These were developed further within the subspace clustering approach, resulting in elegant algorithms and recovery results, such as Elhamifar and Vidal (2009). A related line of work (Livni et al., 2013; Heldt et al., 2009) suggested algorithms that approximate the support of a distribution using an algebraic set. They showed that estimating this algebraic set can be a tractable task but did not devise an algorithm for exploiting unlabeled data.

3. Problem setup

In the standard supervised learning setting a labeled sample of m points is provided. In semi-supervised learning one is additionally provided with a set of unlabeled points. To simplify analysis we assume that the full distribution over X is provided (Section 6 addresses the more realistic version of a finite set of unlabeled points). The corresponding SSL algorithm is defined below. We use binary classification for simplicity, with labels $Y = \{-1, 1\}$.

Definition 1 (SL and SSL) *A semi-supervised learning algorithm B is a function which receives a labeled sample $\{x_i, y_i\}_{i=1}^m$ (where $x_i \in X, y_i \in Y$) and a distribution $P(X)$ over X, and returns a classification rule $h : X \to Y$.*

A supervised learning algorithm A takes only $\{x_i, y_i\}_{i=1}^m$ as input, and returns a classification rule.

Intuitively, SSL is helpful if after having seen the unlabeled data, we can achieve a certain accuracy with much fewer labeled examples than needed without the unlabeled data. To quantify this effect, we next follow Ben-David et al. (2008) and provide several definitions of the relative sample complexity of SL and SSL. Before providing those, we note that a notion of SL-SSL gap cannot be completely distribution free, since the distribution used for SL lower bounds (i.e., the uniform distribution over a shattered set. See Vapnik (1998)) can be used to obtain the same lower bounds for SSL. Thus, Ben-David et al. (2008) define SSL to be effective when there *exists* a distribution such that the sample complexity of SSL is bounded away from that of SL. This is reflected in the two definitions below, which follow Ben-David et al. (2008).

Definition 2 (Sample complexity (SL and SSL)) *For a class $\mathcal{H}$ of hypotheses, the sample complexity of a semi-supervised algorithm B with respect to $P(X,Y)$ (the data generating distribution), is a mapping from $\epsilon > 0$ and $\delta > 0$ to $\mathbb{N}$ such that*

$$m(B,\mathcal{H},P,\epsilon,\delta) = \min\left\{m :\in \mathbb{N} : \mathop{\mathbb{P}}_{S\sim P^m}\left[err^P(B(S,P(X))) - \inf_{h\in\mathcal{H}} err^P(h) > \epsilon\right] < \delta\right\}. \quad (1)$$

Where $P(X)$ is the marginal X distribution of $P(X,Y)$ and $err^P(h)$, is the generalization error of hypothesis h under distribution P.

The sample complexity of a supervised learning algorithm with respect to $P(X,Y)$ is defined similarly, but without the dependence of the algorithm on P.

Given a distribution $P(X)$ over X and a hypothesis h, denote by P_h the extension of P to a distribution over $X\times Y$ that is consistent with h. The following measure for evaluating performance for SSL is similar to that of Ben-David et al. (2008):

Definition 3 *A semi-supervised algorithm B is said to benefit from unlabeled data with respect to the hypothesis class $\mathcal{H}$ and distribution class $\mathcal{P}$ if for every constant c and supervised algorithm A there is a distribution $P\in\mathcal{P}$, ϵ and δ such that:*

$$\sup_{h\in\mathcal{H}} m(A,\mathcal{H},P_h,2\epsilon,\delta) > c\cdot\sup_{h\in\mathcal{H}} m(B,\mathcal{H},P_h,\epsilon,\delta),$$

In other words, an SSL algorithm B is good if it has strictly better sample complexity than any SL algorithm A, as demonstrated by a particular distribution $P(X)$. As mentioned above, it was hypothesized in Ben-David et al. (2008)[2] that no such SSL algorithms exists. However, Darnstädt et al. (2013) provided such an example, in a discrete setting. In what follows we provide a geometric setting, and its analysis.

Our focus is mainly on geometric settings, where a learner wishes to exploit the manifold structure of the data in order to benefit from unlabeled data. Therefore we mainly focus on benefit when the class of *non-discrete* distributions (i.e. distributions where each point in Euclidean space has probability 0) is considered. We will denote the class of non-discrete distributions by $\mathcal{P}_*$.

2. We note that, similar to Darnstädt et al. (2013) we allow, due to technicalities, the SL learner to be less accurate by a factor of 2. For the motivation and discussion we refer the reader to (Darnstädt et al., 2013; Darnstädt, 2015)

3.1. The Hypothesis Class of Algebraic Sets

The hypothesis class we consider has a simple geometric structure, as described next. We first recall the definition of algebraic sets. An algebraic set is a set in $\mathbb{R}^m$ that corresponds to the common zeros of some finite set of multivariate polynomials. Up to some singularities, algebraic sets are in fact manifolds (Hartshorne, 1977). Simple examples of algebraic sets are spheres, cylinders, hyperplanes etc.

We consider the hypothesis class $\mathcal{H}_{\text{alg}}$ corresponding to algebraic sets. In other words each $h \in \mathcal{H}$ is defined by a finite set of polynomials, and $h(x) = 1$ if all polynomials vanish at x. Geometrically, each hypothesis is defined by a specific manifold, and all positively labeled points reside on that manifold. This is defined formally below.

Definition 4 (The hypothesis class of algebraic sets) *A set $V \subset \mathbb{R}^m$ is called an algebraic set if it is the locus of zeros of a collection of multivariate polynomials. The hypothesis class of algebraic sets $\mathcal{H}_{\text{alg}}$ consists of all hypotheses h_V where V is an algebraic set and $h_V(x) = 1$ iff $x \in V$.*

3.2. Main Results

Our main result is that there exists an SSL algorithm for the hypothesis class $\mathcal{H}_{\text{alg}}$, which can benefit from unlabeled data, as defined in Definition 3. The following theorem provides a version of the result for the realizable case.

Theorem 5 *Let $\mathcal{H}_{\text{alg}}$ be the hypothesis class of algebraic sets. Then there is a semi-supervised algorithm that benefits from unlabeled data w.r.t. $\mathcal{H}_{\text{alg}}$ and $\mathcal{P}_*$. Specifically, there is a semi-supervised learner B such that for every constant c, ϵ, δ and supervised learner A, there is a distribution $P \in \mathcal{P}_*$ such that:*

$$\sup_{h \in \mathcal{H}_{\text{alg}}} m(B, \mathcal{H}_{\text{alg}}, P_h, \epsilon, \delta) < \frac{2}{\epsilon} \log \frac{2}{\delta} \tag{2}$$

and

$$\sup_{h \in \mathcal{H}_{\text{alg}}} m(A, \mathcal{H}_{\text{alg}}, P_h, 2\epsilon, \delta) > c\frac{2}{\epsilon} \log \frac{2}{\delta} \tag{3}$$

We can prove a similar result for the agnostic (i.e., non-realizable) case (with ϵ^2 instead of ϵ on the RHS). The techniques are similar, so we focus on the realizable case for simplicity.

3.3. The class $\mathcal{P}_k$ and hypothesis class $\mathcal{G}_d$

The first part of our main theorem states that there exists an SSL algorithm B that can learn with the sample complexity upper bounded in Equation 2. As a first step we will need to define a family of distributions for which our upper bounds apply. For proving the lower bound, it will suffice to use a hypothesis from a restricted hypothesis class $\mathcal{G}_d \subseteq \mathcal{H}_{\text{alg}}$, described in Definition 7.

Our proposed distributions $P(X)$ rely on the existence of a particular type of algebraic sets called *irreducible algebraic sets* (e.g., see Šafarevič, 1994). An algebraic set U is said to be irreducible if it is not a proper union of two different algebraic sets. More formally, for every two algebraic sets M and V such that $U = M \cup V$ we have $U = V$ or $U = M$. Every algebraic set can be decomposed into a finite union of irreducible components, and vice versa the union of

irreducible algebraic set results in an algebraic set. Figure 1(*a*) and Figure 1(*b*) show an example of an algebraic set and its irreducible components.

We next define the class of distributions for which our bounds will be demonstrated.

Definition 6 *The class $\mathcal{P}_k$ of distributions consists of all distributions $P(X)$ such that there are k irreducible algebraic sets $V_1, \ldots, V_k$ with the following property: For every irreducible algebraic set U with $P(U) > 0$ we have that for some $J \subseteq [1, \ldots, k]$ $\cup_{j \in J} V_i \subseteq U$ and further $P(U) = P(\cup_{j \in J} V_i)$. We call $V_1, \ldots, V_k$ the irreducible components of the distribution P.*

The class P_k roughly corresponds to unions of irreducible algebraic sets. The definition might seem restrictive, but it is in fact highly expressive and under very mild conditions on a distribution, it will belong to $\mathcal{P}_k$ for some k. From our perspective, the crucial property of irreducible sets is that the intersection of two distinct irreducible algebraic sets U and V will be either U, V or a manifold of strictly lower dimension. In Figure 1(*c*) we depict the intersection of a hyperplane and a torus which leads to a manifold of dimension 1. Under mild conditions on the distribution P, a manifold of strictly smaller dimension will be a null set and so will have measure zero. Hence, in the example depicted, any irreducible algebraic set U that does not contain the blue torus or the red heart shape, will in fact have measure zero, as required from the definition of $\mathcal{P}_2$.

We next introduce a class of hypotheses which will be used for proving the lower bound in Equation 3.

Definition 7 *The hypothesis class $\mathcal{G}_d$ of polynomial graphs up to degree d over the domain $\mathbb{R}^2$ consists of all hypotheses of the form*

$$h_p(x_1, x_2) = 1 \Leftrightarrow p(x_1) = x_2.$$

for some univariate polynomial p of degree at most d.

The class $\mathcal{G}_d$ is a subset of $\mathcal{H}_{\text{alg}}$.[3] Thus it suffices to prove the bound in Equation 3 by restricting h in $\mathcal{G}_d$, for some d. Also, note that by considering the first two coordinates a Euclidean space $\mathbb{R}^r$, we can naturally embed $\mathcal{G}_d$ in any domain $\mathbb{R}^r$ and our results are therefore not restricted to the plane.

In what follows we prove the upper and lower bounds in Theorem 5. In Section 4, we show that any distribution in $\mathcal{P}_2$ results in the upper bound of Equation 2. Then in Section 5 we show that there exist distributions in $\mathcal{P}_2$ that result in the lower bound of Equation 3. These two facts then imply Theorem 5.

4. Upper Bound for the Semi-supervised Learner

We turn to describing the SSL algorithm B which achieves the upper bound in Equation 2. The algorithm can be explained via the example in Figure 1(*c*), where we assume that the marginal distribution is supported on the torus and heart shapes. This marginal distribution will belong to $\mathcal{P}_2$ (see Definition 6). Now, assume that the learner observes two positive labeled points on each of the yellow circles and one negative point on the heart-shape. There are many $h \in \mathcal{H}_{\text{alg}}$ that agree with the observed labels. For example the hyperplane shown in Figure 1(*c*) also achieves zero training

3. This follows since sets in $\mathcal{G}_d$ correspond to vanishing points of the polynomial $p(x_1) - x_2$, where p is as in the above definition.

error. However, given the known distribution $P(X)$ we realize that the hyperplane will have zero probability (regardless of class) and thus it wouldn't make sense to choose this h. Indeed, the only reasonable choice in this setting is to choose h that corresponds to the blue torus. To summarize, B proceeds as follows: it first decomposes $P(X)$ into irreducible components. Then, upon observing the labeled sample, it considers $h \in \mathcal{H}_{\text{alg}}$ that are unions of these components, and chooses the one with minimum training error.

Figure 1(*c*) also highlights why a supervised learner is bound to have much worse error than B. Since any supervised learning algorithm is oblivious to $P(X)$, it has no way of preferring the blue torus, and will thus unavoidably have high generalization error.

Below we formally state algorithm B and the corresponding guarantee.

Definition 8 ($\mathcal{P}_k$-ERM Algorithm) *A semi-supervised algorithm B is said to be a $\mathcal{P}_k$-ERM if given $P \in \mathcal{P}_k$ with irreducible components $V_1, \ldots, V_k$ and a finite labeled sample $S = \{x_i, y_i\}_{i=1}^m$, the algorithm considers the following class of hypotheses*

$$\mathcal{H}_P = \{h_U : U = \cup_{j \in J} V_j \quad J \subseteq \{1, \ldots, k\}\}$$

and returns h_U that minimizes the empirical error, i.e.

$$B(S, P) = \arg \min_{h_U \in \mathcal{H}_p} |\{h_U(x_i) \neq y_i : (x_i, y_i) \in S\}|$$

Theorem 9 *Let $\mathcal{H} \subseteq \mathcal{H}_{\text{alg}}$ and Let $P \in \mathcal{P}_k$. Any $\mathcal{P}_k$-ERM algorithm B has the following sample complexity bound with respect to P in the realizable case:*

$$\sup_{h \in \mathcal{H}} m(B, \mathcal{H}, \epsilon, \delta, P_h) < \frac{k}{\epsilon} \log \frac{2}{\delta},$$

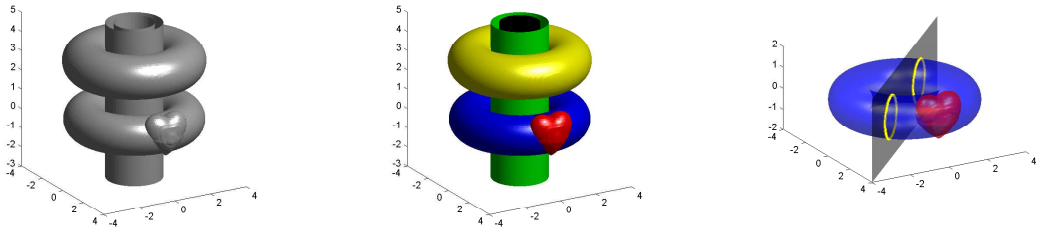
Proof *Let $P \in \mathcal{P}_k$ with corresponding components $V = \cup_{i=1}^k V_i$. For every U, we have that for some $J \subseteq [1, \ldots, k]$, except for a P-null set:*

$$U = \cup_{i \in J} U \cap V_i = \cup_{i \in J} V_i.$$

We conclude that for every U there is a hypothesis in $\mathcal{H}_P$ that minimizes the generalization error over P_U. Note that there are exactly 2^k hypotheses in $\mathcal{H}_P$ and a $\mathcal{P}_k$- ERM algorithm is an ERM algorithm over this finite hypothesis class. Our result now follows from standard sample complexity results for learning finite hypothesis classes. ■

5. Lower bound for Supervised Learner

We now turn to discuss our lower bound guarantees in Equation 3. To prove the result, we need to show that for every supervised learner A and constants ϵ, δ there is a distribution $P \in \mathcal{P}_2$ and a hypothesis $h \in \mathcal{H}_{\text{alg}}$ such that $m(A, \mathcal{H}_{\text{alg}}, P_h, \epsilon, \delta)$ is arbitrarily large. We will show this by first proving a lower bound on the expected error of A, and then use a standard argument to obtain a lower bound on the sample complexity.



(a) An algebraic set (b) The components (c) Intersection of algebraic sets

Figure 1: **(a)** An algebraic set in $\mathbb{R}^3$. It is the union of five irreducible algebraic sets: two cylinders, two tori and a heart shape. **(b)** The five irreducible components. The decomposition is unique and well defined. **(c)** An algebraic set (torus and heart shape) intersecting a distinct irreducible algebraic set (hyperplane). The intersection is a 1-dimensional manifold (two circles) of strictly lower dimension. Under weak assumptions on the distribution, this is a null set.

To illustrate the main challenges let us return to the classical lower bound and see why this proof strategy is not applicable in our setting. For lower bounds based on VC dimension (see for example Devroye et al., 1996) one considers a set of k shattered points, and then invokes a probabilistic argument, where a sample and a hypothesis h are drawn independently. It is then easy to show that without observing $O(k)$ points, the supervised learner will fail in expectation.

Such an argument cannot work in our setting, since the distribution $P(X)$ is fixed and known. Any lower bound that is true for a fixed distribution P, must also apply to a semi-supervised learner. Indeed, for any semi-supervised learner B and distribution P, we can always define a supervised learner $A(S) = B(S, P)$ that has the same guarantees as B w.r.t P. One can indeed verify that a uniform distribution over k points is in $\mathcal{P}_k$ and hence our upper bound for SSL and lower bound for SL would coincide.

In order to truly achieve a lower bound, not only does the hypothesis h need to be chosen randomly, but the marginal distribution $P(X)$ should also be chosen randomly, so as to show that there is at least one *bad* distribution for the supervised learner. Perhaps the simplest way to try and generate such a process is to choose several components (e.g., the five components depicted in Figure 1(*b*)), and then randomly choose two out of these which will be used to define a distribution in $P(X) \in \mathcal{P}_2$ (e.g, corresponding to the uniform distribution over the two components). This simple strategy will however not work. To see why, consider a learner A which "knows" the sampling process. Then once A observes an example on a component, it will identify the component, and will quickly have full information about the components of $P(X)$.[4]

To correct the above strategy, the process as a whole, should be "rich" enough, so that with a limited sample, the learner cannot identify the marginal distribution. A further complication is that h and $P(X)$ cannot be chosen independently, since if h doesn't share components with the irreducible components of P all labels are negative (a case which a supervised learner can learn very quickly).

4. Unless the sample contains a point in the intersection, but these have zero probability.

To summarize, proving the lower bound requires some random process for picking an unlabeled sample and its labels. This process needs to be broken into two parts – first a random algebraic set is chosen, and then labeled data on this surface is generated according to a distribution supported on this surface. This is not trivial in our continuous setting, as it results in certain high dimensional integrals. To evaluate these we use the coarea formula (see below), which essentially allows writing a stochastic process as a double integral, over the image and pre-image of a function F.

5.1. The Coarea Formula

For the following theorem see, for example, in Krantz and Parks (2008) (Corollary 5.2.6):

Theorem 10 *[Coarea Formula] If $F : \mathbb{R}^m \to \mathbb{R}^n$ is Lipschitzian and $m > n$, then for any measurable set K and a measurable function $g : \mathbb{R}^m \to \mathbb{R}$*

$$\int_K g(z) JF(z) d\lambda^m = \int_{\mathbb{R}^n} \left[\int_{F^{-1}(w)\cap K} g(z) d\mathbb{H}^{m-n} \right] d\lambda^n(w) \tag{4}$$

where $\mathbb{H}^{m-n}$ is the $m-n$ dimensional Hausdorff measure in $\mathbb{R}^m$, λ^k is the Lebesgue measure over $\mathbb{R}^k$, $F^{-1}(w)$ is the pre image of w under F and:

$$JF(z) = \sqrt{\det\left(DF(z) \cdot DF(z)^\top\right)}$$

where $DF(z)$ is the Jacobian matrix of F.

The theorem essentially shows how to break down an integral of a function $g(z)$ on $\mathbb{R}^m$ into a double integral over $\mathbb{R}^n$ and $\mathbb{R}^{m-n}$. The double integral first integrates over $w \in \mathbb{R}^n$ and then over the pre image of w under F. The factor JF acts as the appropriate volume element.

We next rewrite the coarea formula in terms of probability measures. Let $F : K \to \mathbb{R}^n$, be a Lipschitz function defined on a subset $K \subset \mathbb{R}^m$. Applying Kirszbraun's Theorem we may assume that F is defined on the whole space and is Lipschitzian and we can apply the coarea formula to F in the domain K.

Let us denote by P_F the probability measure whose density function is given by JF. Namely:

$$P_F(A) = \frac{\int_{K\cap A} JF(z)dz}{\int_K JF(z)dz}$$

P_F induces a distribution on $w \in \mathbb{R}^n$ that by abuse of notation we also denote by P_F. Specifically for every $W \subset \mathbb{R}^n$:

$$P_F(W) = P_F(F^{-1}(W)) = P_F(\{x : F(x) \in W\}).$$

Applying the coarea formula to $g(z) = \chi_{F^{-1}(W)}$ we can write:

$$P_F(W) = \frac{1}{\int_K JF(z)dz} \int_W \left[\int_{F^{-1}(w)\cap K} d\mathbb{H}^{m-n} \right] d\lambda^n(w).$$

Additionally for each w let us denote by $P_F(\cdot|w)$ the probability measure that is supported on $F^{-1}(w) \cap K$ and is given by

$$P_F(A|w) = \frac{\int_{F^{-1}(w)\cap K\cap A} d\mathbb{H}^{m-n}}{\int_{F^{-1}(w)\cap K} d\mathbb{H}^{m-n}}.$$

With these notations we can rewrite the coarea formula as follows

$$\mathop{\mathbb{E}}_{z\sim P_F}[g(z)] = \mathop{\mathbb{E}}_{w\sim P_F}\left[\mathop{\mathbb{E}}_{z\sim P_F(\cdot|w)}[g(z)]\right] \tag{5}$$

5.2. Lower Bound – Main Statement and Proof Sketch

Let $A : ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)) \to \{-1, 1\}$ be some supervised algorithm. Given a labeled sample set $S_n = ((\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n))$ and a test point $(\boldsymbol{x}, y)$ we will denote by $\chi_A(S_n, (\boldsymbol{x}, y))$ the error made by A on the test point when learning from the sample S_n. Namely

$$\chi_A(S_n, (\boldsymbol{x}, y)) = \begin{cases} 1 & A(S_n)[\boldsymbol{x}] \neq y \\ 0 & \text{else} \end{cases}.$$

As stated, using standard techniques we can bound the sample complexity by first bounding the expected error of a supervised learner. Specifically, our lower bound will follow immediately from the following statement:

Theorem 11 *For every learner A and integer N there is a $P \in \mathcal{P}_2$ and a hypothesis $V_1 \in \mathcal{G}_N \subset \mathcal{H}_{\text{alg}}$ such that for every n:*

$$\mathop{\mathbb{E}}_{S_n\sim P_{V_1}}\mathop{\mathbb{E}}_{(\boldsymbol{x},y)\sim P_{V_1}}[\chi_A(S_n, (\boldsymbol{x}, y)] \geq \frac{1}{4}\left(1 - \frac{n^2 - n}{2N}\right)$$

As mentioned earlier, the proof of the theorem is carried through a probabilistic argument, where the main tool is Equation 5. A complete proof is provided in Appendix B. Here we outline the proof (see also Figure 2). As a preliminary step, to apply the coarea formula, we will first need to define a set of functions $F_{\boldsymbol{y}}$ (one for every $\boldsymbol{y} \in \{-1, 1\}^N$) and family of parametrized pairs of irreducible algebraic sets $\{(V_1^w, V_{-1}^w)\}_{w\in\mathbb{R}^N}$ that will satisfy

$$F_{\boldsymbol{y}}^{-1}(\omega) = V_{y_1}^\omega \times \ldots \times V_{y_N}^\omega. \tag{6}$$

Thus, given a sample S_N, if we define $F(S_N) = F_{\boldsymbol{y}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ then the preimage of a pair of algebraic sets is exactly all samples drawn from the pair, consistent with the labeling.

Once this is established, the proof follows a two step argument:

Step 1: We define a random process where we pick random points $\boldsymbol{z}^{(N)} = \{\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N\}$ and a random labeling $y_1, \ldots, y_N$. The points $\boldsymbol{z}^{(N)}$ will actually not be IID, but a bound for the IID case can be derived from this process (see comment following step 2).

We then let $\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})$ measure the expected loss on a random subsample of size $n-1$ and a test point. Specifically:

$$\ell_n(\boldsymbol{z}^{(N)}; \boldsymbol{y}) = \mathop{\mathbb{E}}_{t_1,\ldots,t_n\sim[N]}\left[\chi_A\left(S_{n-1}, (z_{t_1}, y_1)\right)\right], \quad S_{n-1} = ((\boldsymbol{z}_{t_2}, y_{t_2}), \ldots, (\boldsymbol{z}_{t_n}, y_{t_n}),).$$

We next lower bound the expected loss, $\mathbb{E}_{\boldsymbol{y}\sim U}\mathbb{E}_{\boldsymbol{z}^{(N)}\sim P_{F_{\boldsymbol{y}}}}\left[\ell_n(\boldsymbol{z}^{(N)},\boldsymbol{y})\right]$. For the lower bound we exploit the fact that labels and samples were chosen arbitrarily (this is more challenging than the standard supervised lower bound since the distribution of the samples does depend on the $\boldsymbol{y}$. However, by properties of $F_{\boldsymbol{y}}$ we can still prove a lower bound) .

Step 2: We now exploit Equation 5 to show that Step 1 is equivalent to the following process. First, draw ω and $\boldsymbol{y}$ (according to some distribution). Next, draw vectors $\boldsymbol{z}^{(N)}$ such that $F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) = \omega$. Note that Equation 6 states that $F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) = \omega$ implies $\boldsymbol{z}_k \in V^{\omega}_{y_k}$. In other words, the inner process (of picking $\boldsymbol{z}^{(N)}$ from the preimage of ω) is a process where we pick points on algebraic sets: positive from V^{ω}_1 and negatives from V^{ω}_{-1}.

This equivalence then implies (via the standard probabilistic method argument) that there exists a hypothesis $h \in \mathcal{H}_{\text{alg}}$ (namely V^{ω}_1 for some ω) and a distribution $P(X) \in \mathcal{P}_2$ (namely some distribution with support $V^{\omega}_1 \cup V^{\omega}_{-1}$) such that if we learn these with A, we will not be able to generalize.

The above process may generate sample points that are not distributed IID. Since generalization bounds require IID sampling, as a final step we relate the above expected error to that of the corresponding IID sample.

Theorem 11 refers to the expected generalization error over training samples. This can be converted to the result in Equation 3 via standard arguments. See details in section B.4.

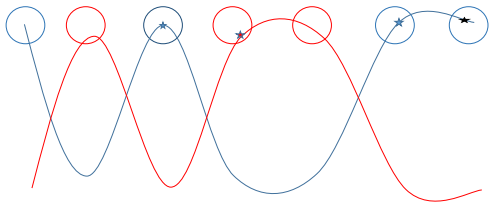


Figure 2: An illustration of the construction used for the lower bound proof. We consider two sampling procedures, and show their equivalence. In the first procedure we assign a random label to each of the N balls in the picture, and then sample N points from the N balls (one from each ball). These points are then assigned the corresponding labels, and the learner A is given a subsample of size n out of these as a training set (these are the blue and red stars in the picture). The resulting hypothesis is applied to another point (the black star in the picture) and the expected loss is measured. This procedure will result in a large error. We next use the coarea formula to show equivalence to a second procedure, which is used to show the lower bound. In the second procedure, two curves are sampled such that they pass through all N balls. One of the curves is assigned label $+1$ and the other -1. Next, n points are sampled from the balls (on the curves), and A learns from those. By showing that the two procedures are equivalent, we can conclude that there exist an input distribution $P(X) \in \mathcal{P}_2$ and a hypothesis $h \in \mathcal{H}_{\text{alg}}$ which cannot be learned by A.

6. A Practical Algorithm

The SSL algorithm B described earlier has two main steps: find the irreducible components of $P(X)$ and find a hypothesis minimizing error on those. To turn it into a practical algorithm two

issues need to be addressed. First, our analysis assumed the distribution $P(X)$ is given explicitly , but in practice we only learn about it from observing a set of unlabelled points sampled from $P(X)$. Second, even if we know $P(X)$, factoring a distribution into irreducible components is notoriously difficult (Wang, 1992). Below we argue that in practice both problems can be circumvented, under certain assumptions, by employing ideas from subspace clustering (e.g., see Vidal et al., 2005; Vidal, 2010; Parsons et al., 2004; Elhamifar and Vidal, 2009).

Subspace clustering is defined as the problem of decomposing a set of points into k disjoint linear subspaces. The problem is generally hard, even assuming the data indeed corresponds to such k subspaces. However, under certain assumptions there are efficient algorithms with statistical recovery guarantees. For example, the semi-random subspace clustering model in Soltanolkotabi et al. (2012) assumes the following generative process: k subspaces $S_1, \ldots, S_k$ are chosen and points are sampled IID from the union of these linear subspaces. The aim of a subspace clustering algorithm is to recover the subspaces $S_1, \ldots, S_k$. Under a uniform distribution assumption, Park et al. (2014) then provide an efficient algorithm to recover the underlying subspaces. See also Soltanolkotabi et al. (2014) for another subspace clustering algorithm with recovery guarantees.

To relate the problem of subspace clustering to our model, let us start with the simple case where $P(X)$ is a distribution in $\mathcal{P}_k$ and the associated components $V_1 \cup V_2 \cup \ldots \cup V_k$ are all linear subspaces. By definition a sample drawn from P will result in an IID randomly generated sample as required in the semi-random subspace clustering model. To apply a $\mathcal{P}_k$–ERM algorithm, we only need to recover the irreducible components $V_1, \ldots, V_k$, which is exactly the task of a subspace clustering algorithm. Thus, subspace clustering is in fact a particular case of our setting, where one assumes that the algebraic sets are linear subspaces. A subspace clustering algorithm simultaneously addresses the two practical issues mentioned above by directly factoring $P(X)$ as required by our SSL algorithm.

Although subspace clustering is applied to linear subspaces, it may be readily applied to algebraic sets via the standard approach of expanding the space of features with all monomials of the degree of the polynomials underlying the algebraic sets (Vidal et al., 2005). For example the algebraic set corresponding to a circle $x_1^2 + x_2^2 - 1 = 0$ is a linear subspace in the feature space $\left[x_1, x_2, x_1 x_2, x_1^2, x_2^2, 1\right]$. In the same way, a union of algebraic sets will correspond to a union of linear subspaces in the expanded feature space. The task of decomposing the unlabeled data into algebraic sets is thus reduced to subspace clustering in the new feature space of dimension m^d where d is the total degree of generating polynomials. Specifically, for every d we let $\Omega_d := \{\omega \in \mathbb{N}^m : \sum \omega_i \leq d\}$ and let $\phi_d(\boldsymbol{x}) \in \mathbb{R}^{\Omega_d}$ be the embedding into monomial space, i.e. $(\phi_d(\boldsymbol{x}))_\omega := \prod \boldsymbol{x}_i^{\omega_i}$. Then, via the embedding ϕ_d, the task of finding algebraic sets $V_1, \ldots, V_k$ with generators of total degree d is reduced to the task of subspace clustering in dimension m^d.

The above monomial expansion seems to require exponential time and space. However, a standard kernel trick can be employed, resulting in runtime polynomial in d. This follows from the fact that subspace clustering algorithms such as Soltanolkotabi et al. (2014) rely only on dot products and can thus be "kernelized" avoiding the need for explicit calculation in monomial space..

To summarize: the practical version of our SSL algorithm is to first run a kernelized subspace clustering algorithm on the unlabeled points. This will find the irreducible components of $P(X)$. Next, we just need to assign the labeled points to these components and run the $\mathcal{P}_k$–ERM. This method can always be used, but we note that it will provably recover the underlying components only if $P(X)$ satisfies the conditions in the corresponding subspace clustering algorithms.

7. Discussion

We present a learning setting where classes correspond to algebraic sets, and SSL is arbitrarily better than SL.

As mentioned earlier, one of the main approaches to analyzing SSL is via the notion of compatibility. In Balcan and Blum (2010) a notion of compatibility $\chi : \mathcal{H} \times \mathcal{X} \to \{0, 1\}$ is defined as a function from a hypothesis and a sample point to $\{0, 1\}$. Denoting the expected compatibility by $\chi(f, P) = \mathbb{E}_{x\sim P}[\chi(f, x)]$, the unlabeled error rate of f with P is defined as $1 - \chi(f, P)$. As an example, a classifier f may have high compatibility with P if its separator passes through low density regions of P.

Though in our model we do not consider a compatibility notion, our results can be cast to this model by considering a degenerate compatibility notion $\chi(f, x) = 1$. Within this framework, our work demonstrates a setting where the learning rate of the SSL is arbitrarily better than that of any SL learner. Balcan and Blum (2010) studied two types of bounds in the context of SSL. The first is *uniform convergence* bounds which rely on the convergence of the empirical risk to the generalization error *uniformly* over all concepts that are compatible. The second is ϵ*–cover* bounds which rely on replacing the original class with a smaller class that can ϵ–approximate any compatible concept (in expectation, w.r.t the marginal distribution). Our analysi relies on an ϵ–cover bound (as demonstrated, for the given marginal distribution: every concept is equivalent to some concept in the smaller class $H_{\mathcal{P}}$).

Another interesting feature of our result is its implication to VC bounded classes. Our result demonstrates a class with unbounded VC dimension for which the benefit can be arbitrarily large. It is known that for VC bounded classes the benefit cannot be arbitrarily large. Specifically, for VC bounded classes Darnstädt et al. (2013) showed that the benefit can be *at most* order of $O(\text{VC–dim})$. Whether this bound is tight was left as an unresolved question.

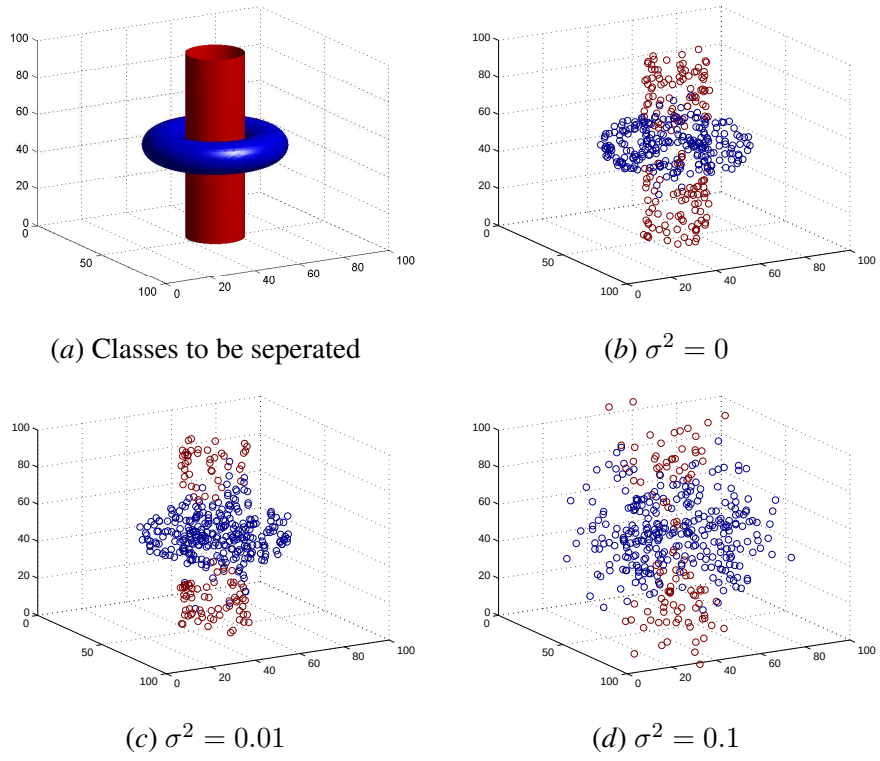
Our construction in Theorem 11 shows that for a fixed $\epsilon > 0$, say, $\epsilon_0 = \frac{1}{8}$, we can choose a concept from a class of VC-dimension N, for which any SL learner needs to observe at least $O(\sqrt{N})$ examples in order to achieve ϵ_0 accuracy. In contrast the SSL learner needs to observe $O(1)$ examples. This means that we can have benefit of *at least* $\Omega(\sqrt{\text{VC-dim}})$, for fixed ϵ_0. Previous construction by Darnstadt et al. relied on Borell-Cantelli Lemma and it is not immediate to relate the result to a finite VC class. It would be interesting to further close the asymptotic gap for finite VC classes.

8. Synthetic Experiment

In the experiments below we employ the subspace clustering method of Soltanolkotabi et al. (2014). We consider a toy example involving two algebraic sets: a torus and a cylinder (see Figure 3(*a*)). Both sets are defined by polynomial equations of total degree 4. Subspace clustering in monomial space was used to divide the labeled data into two subspaces and extract a basis for vanishing polynomials of total degree d.

In the case of $P \in \mathcal{P}_2$ there is a particularly simple method to learn $\mathcal{H}_P$: Let $V_1 = \{\boldsymbol{x} : p_1(\boldsymbol{x}), \ldots, p_{\ell_1}(\boldsymbol{x}) = 0\}$ and $V_2 = \{\boldsymbol{x} : q_1(\boldsymbol{x}), \ldots, q_{\ell_2}(\boldsymbol{x}) = 0\}$ be the algebraic support. Then the two classes are separable in the feature space:

$$\boldsymbol{x} \to \left(p_1^2(\boldsymbol{x}), \ldots p_{\ell_1}^2(\boldsymbol{x}), q_1^2(\boldsymbol{x}), \ldots q_{\ell_2}^2(\boldsymbol{x})\right) \tag{7}$$



(a) Classes to be seperated (b) $\sigma^2 = 0$

(c) $\sigma^2 = 0.01$ (d) $\sigma^2 = 0.1$

Figure 3: The Subspace clustering performed by the Robust SSC procedure, with different noise levels. Figure 3(a) illustrates the two real components from which the data was generated.

Indeed by taking

$$\sum q_i^2(\boldsymbol{x}) - \sum p_i^2(\boldsymbol{x}) \tag{8}$$

for any $\boldsymbol{x} \in V_1$ we have that all p_i's are zero, and there exists a j such that $q_j(\boldsymbol{x}) \neq 0$. Hence, the overall sum in Equation 8 will be positive on V_1 and negative on V_{-1}.

More generally, the class $\mathcal{H}_P$ can be learned by learning a linear classifier over the feature space in Equation 7. Thus, after extracting the vanishing polynomials, we embedded as suggested in the corresponding feature space and performed linear classification.

Our supervised learner for comparison was chosen to be an SVM with polynomial kernels up to degree 8 (as it will have the same expressive power as our SSL learner). To verify the robustness to noise of our approach, we also tested the effect of adding noise. The learning curve is shown in Figure 4. It can be seen that our SSL algorithm indeed learns with fewer samples, in agreement with Theorem 5. In the high noise case ($\sigma^2 = 0.1$) the subspace clustering algorithm often fails, and as a result the gap in sample complexity is smaller.

Acknowledgements The authors would like to thank the anonymous reviewers for their helpful comments. This work was supported by the Eric and Wendy Schmidt Fund for Strategic Innovation, a Google Research Award, the Blavatnik Computer Science Research Fund, and the Intel Collaborative Research Institute for Computational Intelligence (ICRI-CI).

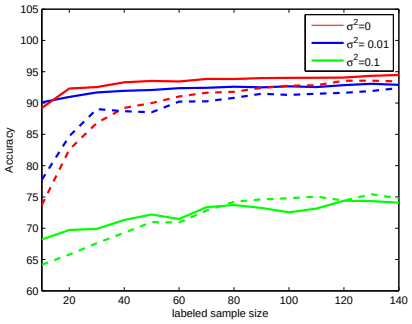


Figure 4: Supervised Learner vs. Semi Supervised Learner. Solid lines are SSL, dashed lines are SL. We measure accuracy as a function of labeled sample size. The SSL algorithm was based on subspace clustering, as described in Section 8.

References

Yasemin Altun, David McAllester, and Mikhail Belkin. Maximum margin semi-supervised learning for structured variables. *Advances in neural information processing systems*, 18:33, 2006.

Martin Azizyan, Aarti Singh, and Larry Wasserman. Density-sensitive semisupervised inference. *The Annals of Statistics*, 41(2):751–771, 2013.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):19, 2010.

Mikhail Belkin and Partha Niyogi. Semi-supervised learning on Riemannian manifolds. *Machine Learning*, 56(1):209–239, 2004.

Shai Ben-David and Ruth Urner. The sample complexity of agnostic learning under deterministic labels. In *Proceedings of The 27th Conference on Learning Theory*, pages 527–542, 2014.

Shai Ben-David, Tyler Lu, and Dávid Pál. Does unlabeled data provably help? worst-case analysis of the sample complexity of semi-supervised learning. In *Proceedings of the 21st Annual Conference on Learning Theory*, pages 33–44, 2008.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. *Proc. 18th International Conf. on Machine Learning*, 2001.

Vittorio Castelli and Thomas M Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

Olivier Chapelle, Bernhard Schölkopf, Alexander Zien, et al. *Semi-supervised learning*. MIT press Cambridge, 2006.

Fabio Gagliardi Cozman, Ira Cohen, Marcelo Cesar Cirelo, et al. Semi-supervised learning of mixture models. In *Proceedings of the Twentieth International Conference on Machine Learning*, pages 99–106, 2003.

Malte Darnstädt. *An investigation on the power of unlabeled data.* PhD thesis, Ruhr University Bochum, 2015.

Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Unlabeled data does provably help. In *STACS*, pages 185–196, 2013.

L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Thoery of Pattern Recognition*. Springer, 1996.

Mathias Drton, Bernd Sturmfels, and Seth Sullivant. *Lectures on algebraic statistics*, volume 39. Springer Science & Business Media, 2008.

Ehsan Elhamifar and René Vidal. Sparse subspace clustering. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2790–2797. IEEE, 2009.

Paolo Gibilisco. *Algebraic and geometric methods in statistics*. Cambridge University Press, 2010.

Robin Hartshorne. *Algebraic geometry*, volume 52. Springer Science & Business Media, 1977.

Daniel Heldt, Martin Kreuzer, Sebastian Pokutta, and Hennie Poulisse. Approximate computation of zero-dimensional polynomial ideals. *J. Symb. Comput.*, 44(11):1566–1591, 2009.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *Proceedings of the Second International Conference on Learning Representations (ICLR 2014)*, April 2014.

Franz J Király, Paul Von Bünau, Jan Saputra Müller, Duncan AJ Blythe, Frank C Meinecke, and Klaus-Robert Müller. Regression for sets of polynomial equations. In *AISTATS*, pages 628–637, 2012.

Steven G Krantz and Harold R Parks. *Geometric integration theory*. Springer Science & Business Media, 2008.

Jason Lafferty, Xiaojin Zhu, and Yan Liu. Kernel conditional random fields: representation and clique selection. In *Proceedings of the twenty-first international conference on Machine learning*, page 64. ACM, 2004.

John D. Lafferty and Larry A. Wasserman. Statistical analysis of semi-supervised regression. In *Advances in Neural Information Processing Systems 20*, pages 801–808, 2007.

Roi Livni, David Lehavi, Sagi Schein, Hila Nachliely, Shai Shalev-Shwartz, and Amir Globerson. Vanishing component analysis. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 597–605, 2013.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Computer Science Dept., University of Chicago, Tech. Rep. TR-2008-01*, 2008.

Dohyung Park, Constantine Caramanis, and Sujay Sanghavi. Greedy subspace clustering. In *Advances in Neural Information Processing Systems*, pages 2753–2761, 2014.

Lance Parsons, Ehtesham Haque, and Huan Liu. Subspace clustering for high dimensional data: a review. *ACM SIGKDD Explorations Newsletter*, 6(1):90–105, 2004.

Salah Rifai, Yann Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. The manifold tangent classifier. In *NIPS*, volume 271, page 523, 2011.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *arXiv preprint math/0604233*, 2006.

Igor Rostislavovič Šafarevič. *Basic algebraic geometry*, volume 1. Springer, 1994.

Aarti Singh, Robert Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in neural information processing systems*, pages 1513–1520, 2009.

Mahdi Soltanolkotabi, Emmanuel J Candes, et al. A geometric analysis of subspace clustering with outliers. *The Annals of Statistics*, 40(4):2195–2238, 2012.

Mahdi Soltanolkotabi, Ehsan Elhamifar, Emmanuel J Candes, et al. Robust subspace clustering. *The Annals of Statistics*, 42(2):669–699, 2014.

Vladimir Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

René Vidal. A tutorial on subspace clustering. *IEEE Signal Processing Magazine*, 28(2):52–68, 2010.

René Vidal, Yi Ma, and Shankar Sastry. Generalized principal component analysis (gpca). *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(12):1945–1959, 2005.

Dongming Wang. Irreducible decomposition of algebraic varieties via characteristics sets and gröbner bases. *Computer Aided Geometric Design*, 9(6):471–484, 1992.

Sumio Watanabe. *Algebraic geometry and statistical learning theory*, volume 25. Cambridge University Press, 2009.

Hassler Whitney. Elementary structure of real algebraic varieties. *The Annals of Mathematics*, 66 (3):545–556, 1957.

Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33nd International Conference on Machine Learning (ICML)*, pages 40–48, 2016.

Xiaojin Zhu. Semi-supervised learning literature survey. *Computer Science, University of Wisconsin-Madison*, 2:3, 2006.

Appendix A. Preliminaries

In this section we develop the main properties of algebraic sets that will be needed in order to extract lower bounds. We first recall the definitions of irreducible algebraic sets and irreducible components:

Definition 12 (irreducible algebraic set) *An algebraic set V is said to be irreducible if for any decomposition $V = U \cup M$ where U and M are also algebraic sets, we must have $U = V$ or $V = M$.*

Definition 13 (irreducible component) *An irreducible algebraic set V is called an irreducible component of an algebraic set U, if $U = V$ or $U = V \cup M$ where M is an algebraic set distinct from U.*

A.1. Dimension of Algebraic Sets

We begin by defining the dimension of an algebraic set which is roughly the manifold's dimension:

Definition 14 (Dimension of the tangent space) *For an algebraic set $V \subseteq \mathbb{R}^m$ and $p \in V$:*

$$d = \textit{rank}_p(V) \tag{9}$$

is the maximal number of polynomials $f_1, \ldots, f_d$ such that $f_1, \ldots, f_d$ vanish on V (i.e for all $v \in V$ and $i = 1 \ldots, d$ we have $f_i(v) = 0$) such that the Jacobian matrix $\left(\frac{\partial f_i(p)}{\partial x_j}\right)_{i,j}$ has rank d. The number $m - \textit{rank}_p(V)$ is called the dimension of the tangent space at at a point p of V.

Definition 15 (Dimension of an algebraic set) *The dimension of an irreducible algebraic set V is*

$$\dim V = m - \max\{\textit{rank}_p(V) : p \in V\}.$$

The dimension of an algebraic set is the maximum dimension of its irreducible components.

In general, for any differentiable function $f : \mathbb{R}^m \to \mathbb{R}^n$ with $m \geq n$, at any value r, if $f(x) = r$ and the Jacobian matrix of f has full rank, then locally at x the preimage $f^{-1}(r)$ is a $m-n$ dimensional manifold. Recall that V is the preimage of 0 under a mapping $[f_1, \ldots, f_d]$ of polynomials, and it is a manifold in $\mathbb{R}^m$. At any non-singular point, the dimension of the tangent plane at point p of V coincide with the dimension of V, if V is irreducible.

A.2. Real Algebraic Sets and Their Structure

The following facts follow from Lemmas 6,7,8 in Whitney (1957) and the correctness of these statements in the complex case (see for example Šafarevič (1994)).

Fact A.1 *If V is an irreducible algebraic set and U is a proper sub algebraic set (i.e. U is an algebraic set and $U \subsetneq V$) then U has a strictly lower dimension.*

Fact A.2 *Every strictly descending chain of algebraic sets is finite i.e. if*

$$U_1 \supseteq U_2 \supseteq U_3 \supseteq \ldots.$$

Then for some m we have

$$U_m = U_{m+1} = U_{m+2} = \ldots.$$

Also the following fact can be found for example in Šafarevič (1994)

Fact A.3 *If U and V are two algebraic sets then also $U \cup V$ is an algebraic set and, $U \cap V$ is also an algebraic set. In fact, even infinite intersection of algebraic sets results in an algebraic set*

Corollary 16 *Let P be a distribution over $\mathbb{R}^m$. Then there exists a unique minimal (with respect to the inclusion ordering) algebraic set V such that $P(V) = 1$.*

Proof Start with $V_1 = \mathbb{R}^m$. Assume there is an algebraic set $V_2 \subsetneq V_1$ such that $P(V_2) = 1$. Define inductively for each m an algebraic set $V_{m+1} \subsetneq V_m$ such that $P(V_{m+1}) = 1$. Since every such chain needs to be finite (by Fact A.2 above), we must end with some m such that $P(V_m) = 1$ and for every proper sub algebraic set $P(V_{m+1}) \neq 1$.

As for uniqueness, let U and V be two minimal algebraic sets with $P(U) = P(V) = 1$, then $P(U \cap V) = 1$. But $U \cap V$ is also an algebraic set. By minimality $U \cap V$ cannot be a proper subset of V and we get $U = V$. ■

Definition 17 (Algebraic support) *Let P be a probability measure, and V the minimal algebraic set that supports P. We call V the algebraic support of P.*

Definition 18 (Non degenerate distributions) *Let P be a probability measure, and V its algebraic support. P is said to be non-degenerate if for every algebraic set U, if $P(U) > 0$ then $\dim U \geq \dim V$.*

Another interesting corollary of Fact A.2 is that we can always decompose an algebraic set into finite irreducible components. Indeed if V is not irreducible we can write it as $V = V_1 \cup M_1$. If M_1 is irreducible, we are finished, if not then we can decompose $M_1 = V_2 \cup M_2$. Again we get a strictly decreasing chain $V \supseteq M \supseteq M_1 \supseteq M_2 \supseteq \ldots$ since this process must be finite we can get a finite decomposition. We will see that semi-supervised learnability of an algebraic set is a function of the number of irreducible components of the algebraic support of P.

Theorem 19 *Let P be a non degenerate distribution with algebraic support $V = \cup_{i=1}^k V_i$, where V_i are the irreducible components of V. then $P \in \mathcal{P}_k$.*

Proof Let H be an algebraic set with $P(H) > 0$ since $P(V) = 1$ we can assume that $H = H \cap V = \cup_{i=1}^k H \cap V_i := \cup_{i=1}^k H_i$ For each i whenever $\dim(H_i) < \dim(V_i)$, since $\dim(V_i) < \dim(V)$ we have that $P(H_i) = 0$. Let $J \subseteq [1, \ldots, k]$ be the indices of all H_j's such that $\dim(H_j) = \dim(V_j)$. But since V_j are irreducbile, if $\dim(H_j) = \dim(V_j)$ then $H_j = V_j$. It follows that except for maybe a null set $H = \cup_{j \in J} V_j$. ■

Appendix B. Proof for Theorem 11

We begin along the lines of Section 5.2. The first objective is to construct a function F that "chooses" our support.

B.1. Defining F

In this section we prove the following Lemma:

Lemma 20 *For every N and m there exists a set $\mathcal{B} \subset \mathbb{R}^{N\times m}$ of the form:*

$$\mathcal{B} = B_1 \times B_2 \times \ldots \times B_N \subseteq \mathbb{R}^m \times \mathbb{R}^m \ldots \times \mathbb{R}^m,$$

a family of distinct pairs irreducible algebraic sets $\{V_1^\omega, V_{-1}^\omega\}_{\omega\in\mathbb{R}^N}$ and a family of functions $\{F_{\boldsymbol{y}}\}_{\boldsymbol{y}\subseteq\{-1,1\}^N}$, $F_{\boldsymbol{y}} : \mathcal{B} \to \mathbb{R}^N$ such that:

1. *For every ω:*
$$F_{\boldsymbol{y}}^{-1}(\omega) = \left(V_{y_1}^\omega \times V_{y_2}^\omega \times \ldots \times V_{y_N}^\omega\right) \cap \mathcal{B}.$$
2. *For every $\boldsymbol{z}^{(N)} \in \mathcal{B}$:*
$$\min_{\boldsymbol{y}} JF_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) \geq \frac{1}{2}\max_{\boldsymbol{y}} JF_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}).$$

Further $V_1^\omega \in \mathcal{G}_N$ for every ω.

We will need the following Lemma whose proof is somewhat technical and we defer to Section B.7.

Lemma 21 *For every $\alpha \in \mathbb{R}^k$ and $\beta \in \mathbb{R}$ with $\beta, \alpha_i \neq 0$ the set of real zeros of the following polynomial is an irreducible algebraic set:*

$$p(t_1, t_2) = \beta t_2 - \sum_{i=1}^{k} \alpha_i t_1^i.$$

Proof [proof of Lemma 20] We define the function $F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)})$ that is parametrized by a labeling vector. First, we define a matrix $\mathcal{M}(\boldsymbol{z}^{(N)}) \in M_{N\times N}$ such that

$$\left(\mathcal{M}(\boldsymbol{z}^{(N)})\right)_{i,j} = \boldsymbol{z}_{i,1}^{j-1},$$

where $\boldsymbol{z}_{i,1}$ is the first coordinate of sample number i. Next we define a vector $\boldsymbol{v}(\boldsymbol{z}^{(N)}, \boldsymbol{y})$ that is dependent on both sample and labels.

$$\boldsymbol{v}(\boldsymbol{z}^{(N)}, \boldsymbol{y}) = \begin{bmatrix} y_1\boldsymbol{z}_{1,2} \\ \ldots \\ y_n\boldsymbol{z}_{n,2} \end{bmatrix}$$

where as before $\boldsymbol{z}_{i,2}$ is the second coordinate of the i-th example. Finally, we let $F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) \in \mathbb{R}^N$ be the solution of the linear equation:

$$\mathcal{M}(\boldsymbol{z}^{(N)})\boldsymbol{\alpha} = -\boldsymbol{v}(\boldsymbol{z}^{(N)}, \boldsymbol{y})$$

Namely,

$$F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) = -\mathcal{M}^{-1}(\boldsymbol{z}^{(N)})\boldsymbol{v}(\boldsymbol{z}^{(N)}, \boldsymbol{y}) \tag{10}$$

For every $\omega = \omega_1, \ldots, \omega_N$ define for $e \in \{-1, 1\}$

$$V_e^\omega = \{\boldsymbol{x} \in \mathbb{R}^m : ex_2 - \sum_{k=1}^{N} \omega_k {x_1}^{k-1} = 0\}$$

We begin by constructing an open set $\mathcal{A}$ such that:

$$F_{\boldsymbol{y}}^{-1}(\omega) \cap \mathcal{A} = V_{y_1}^{\omega} \times V_{y_2}^{\omega}, \ldots, V_{y_{n-1}}^{\omega} \times V_{y_n}^{\omega} \cap \mathcal{A}. \tag{11}$$

We then refine it to have that 2 holds too.

By Lemma 21 the set $\mathcal{U} = \{\omega : V_e^{\omega} \text{ is an irreducible algebraic set}\}$ is an open set. Further, since $F_{\boldsymbol{y}}$ are continuous we have that $F_{\boldsymbol{y}}^{-1}(\mathcal{U})$ is also open for every $\boldsymbol{y}$. Also note that for every $\boldsymbol{z}^{(N)}$ such that $\boldsymbol{z}_{i,1} \neq \boldsymbol{z}_{j,1}$ for all $i \neq j$, we have that $M(\boldsymbol{z}^{(N)})$ is non-singular[5]. Therefor if we let

$$\mathcal{A} = \cap_{\boldsymbol{y}} F^{-1}(\mathcal{U}) \cap \{\boldsymbol{z}^{(N)} : M(\boldsymbol{z}^{(N)}) \text{ is non singular}\}$$

then $\mathcal{A}$ is an open set.

Next, by definition of F we have that Equation 11 holds. Indeed, for any $\boldsymbol{x}^{(N)}$ in the mentioned product, we know that ω is a solution to the linear equation defined by $\mathcal{M}$ and $\boldsymbol{v}$, hence $F_{\boldsymbol{y}}(\boldsymbol{x}^{(N)}) = \omega$, similarly if $F_{\boldsymbol{y}}(\boldsymbol{x}^{(N)}) = \omega$ we have the reverse implication.

Next, we turn to define a set $\mathcal{B}$ in the domain of $F_{\boldsymbol{y}}$ for which item 2 will hold if we restrict $F_{\boldsymbol{y}}$ to $\mathcal{B}$. Our claim will follow from the following equation whose proof is somewhat technical and is deferred to Section B.7

$$JF_{\boldsymbol{y}}(\boldsymbol{z}^{(N)}) = \frac{1}{\det \mathcal{M}(\boldsymbol{z}^{(N)})} \prod_{k=1}^{N} \sqrt{\left(\sum_{j=1}^{N} j\omega_j \boldsymbol{z}_{k,2}{}^{j-1}\right)^2 + 1}. \tag{12}$$

By the adjugate formula we have that: $\omega = \frac{1}{\det \mathcal{M}(\boldsymbol{z}^{(N)})} \text{adj} M(\boldsymbol{z}^{(N)}) \boldsymbol{v}[\boldsymbol{z}^{(N)}, \boldsymbol{y}]$ (where adj$M$ is the adjugate matrix of M). The main observation is that $JF_{\boldsymbol{y}}(\boldsymbol{z}^{(N)})$ is continuous in any environment where $\det \mathcal{M}(\boldsymbol{z}^{(N)}) \neq 0$, and that if $\boldsymbol{z}_{i,2} = 0$ for all i we have

$$\frac{JF_{\boldsymbol{y}_1}(\boldsymbol{z}^{(N)})}{JF_{\boldsymbol{y}_2}(\boldsymbol{z}^{(N)})} = 1$$

Finally, by continuity we can pick small balls around some $\boldsymbol{z}_{i,2} \neq 0$ but close to 0, contained in $\mathcal{A}$ such that for every $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$:

$$\frac{JF_{\boldsymbol{y}_1}(\boldsymbol{z}^{(N)})}{JF_{\boldsymbol{y}_2}(\boldsymbol{z}^{(N)})} \geq \frac{1}{2}.$$

Since $\mathcal{B} \subseteq \mathcal{A}$ we have $\mathcal{B} \cap \mathcal{A} = \mathcal{B}$ and by Equation 11:

$$F_{\boldsymbol{y}}^{-1}(\omega) \cap \mathcal{B} = V_{y_1}^{\omega} \times V_{y_2}^{\omega}, \ldots, V_{y_{n-1}}^{\omega} \times V_{y_n}^{\omega} \cap \mathcal{B}.$$

Finally note that since for every $\boldsymbol{z}^{(N)} \in \mathcal{B}$ we have $\boldsymbol{z}_{i,2} \neq 0$ we must have that for $\omega = F_{\boldsymbol{y}}(\boldsymbol{z}^{(N)})$ implies, $V_{-1}^{\omega} \neq V_1^{\omega}$ (Because $V_1^{\omega} = V_{-1}^{\omega}$ implies by definition that for every $\boldsymbol{x} \in V_1^{\omega}$, $x_2 = 0$). Hence, as promised V_1^{ω} and V_{-1}^{ω} are distinct. ■

Throughout the next sections we fix N, the mappings $F_{\boldsymbol{y}}$, the set $\mathcal{B}$ and the family of irreducible sets $\{(V_1^{\omega}, V_{-1}^{\omega})\}_{\omega \in \mathbb{R}^N}$.

5. Note that $M(\boldsymbol{z}^{(N)})$ is a Vandermonde matrix, hence unless two coordinates are equal it will be non-singular

B.2. Defining a random process with lower bound on expected error

First, given a fixed sample $\boldsymbol{z}^{(N)} = \boldsymbol{z}_1, \ldots, \boldsymbol{z}_N$ and labeling $\boldsymbol{y}$ let us define the following random process: we randomly pick n indices $t_1, \ldots, t_n$ and consider the sample set of size $n-1$ $S_{n-1}(\boldsymbol{z}^{(N)}, \boldsymbol{t}, \hat{\boldsymbol{y}}) = (\boldsymbol{z}_{t_2}, y_{t_2}), (\boldsymbol{z}_{t_3}, y_{t_3}), \ldots (\boldsymbol{z}_{t_n}, y_{t_n})$ and a test point $(\boldsymbol{z}_{t_1}, y_{t_1})$. We let:

$$\ell_n(\boldsymbol{z}^{(N)}; \boldsymbol{y}) = \mathbb{E}_{\boldsymbol{t}} \left(\chi_A \left(S_{n-1}(\boldsymbol{z}^{(N)}, \boldsymbol{t}, \hat{\boldsymbol{y}}), (\boldsymbol{z}_{t_1}, \hat{y}_1) \right) \right).$$

In this section we prove the following result:

Lemma 22 *Fix N and let $\mathcal{B}$, $\{V_1^\omega, V_{-1}^\omega\}_{\omega \in \mathbb{R}^N}$, $\{F_{\boldsymbol{y}}\}_{\boldsymbol{y} \in \{-1,1\}^N}$ be as in Lemma 20.*

For every A and $n < N$ there are ω and $\boldsymbol{y}$ such that: If we consider a random process where we randomly pick a sample $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N)$ according to the distribution $P_{F_{\boldsymbol{y}}}(\cdot|\omega)$ then

$$\mathbb{E}\left[\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})\right] \geq \frac{1}{4}$$

Proof The proof is via a probabilistic argument, we begin by lower bounding the expectation w.r.t a distribution that picks labeling randomly and uniformly. We claim that

$$\mathbb{E}_{\boldsymbol{y}} \mathbb{E}_{\boldsymbol{z}^{(N)} \sim P_{F_{\boldsymbol{y}}}} \left[\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})\right] \geq \frac{1}{4} \tag{13}$$

Indeed, fix $\boldsymbol{z}^{(N)}$, $\boldsymbol{t}$ and $y_2, \ldots, y_n$.

$$\frac{1}{2} \sum_{y_1 \in \{-1,1\}} \left(\chi_A(S_{n-1}, (\boldsymbol{z}_{t_1}, y_{t_1})) JF_{\hat{\boldsymbol{y}}}(\boldsymbol{z}^{(N)}) \right) \geq \frac{1}{2} \min_{y_1} JF_{\hat{\boldsymbol{y}}}(\boldsymbol{z}^{(N)}) \geq \tag{14}$$

$$\frac{1}{4} \left(\min_{y_1} JF_{\hat{\boldsymbol{y}}}(\boldsymbol{z}^{(N)}) + \min_{y_1} JF_{\hat{\boldsymbol{y}}}(\boldsymbol{z}^{(N)}) \right) \geq \frac{1}{4} \left(\frac{1}{2} JF_{1,y_2,\ldots,y_n}(\boldsymbol{z}^{(N)}) + \frac{1}{2} JF_{-1,y_2,\ldots,y_n}(\boldsymbol{z}^{(N)}) \right)$$

$$= \frac{1}{4} \mathbb{E}_{y_1} \left(JF_{\hat{\boldsymbol{y}}}(\boldsymbol{z}^{(N)}) \right).$$

Since this holds for every $\boldsymbol{z}^{(N)}$,$\boldsymbol{t}$ and $y_2, \ldots, y_n$, by integrating on both sides, the statement follows from the definition of $P_{F_{\boldsymbol{y}}}$.

It follows that for some fixed sequence of labeling:

$$\mathbb{E}_{\boldsymbol{z}^{(N)} \sim P_{F_{\boldsymbol{y}}}} \left[\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})\right] \geq \frac{1}{4}$$

The statement in now an immediate corollary of Equation 5. Indeed we have that there is $\omega \in \mathbb{R}^N$ such that:

$$\mathbb{E}_{\boldsymbol{z}^{(N)} \sim P_{F_{\boldsymbol{y}}}(\cdot|\omega)} \left[\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})\right] \geq \frac{1}{4}$$

■

Let us return to the random process depicted in Lemma 22: We pick representatives $\boldsymbol{z}_i$ from each ball B_i, then we randomly pick n representatives and measure the expected error when the learner observes the last $n-1$ representatives and is tested on the first representative. We've produced a lower bound on the expected error when samples are drawn according to this process. We now relate this expected error to the expected error when sample is picked IID on the same support, which will conclude the proof.

B.3. Lower bound for IID sample

So far we have shown that under a certain distribution on the sample, which is not necessarily IID, we can lower bound the expected error, we now wish to relate this to a process where we pick the sample IID. This section is dedicated to proof of the following statement:

Theorem 23 *For every supervised algorithm A and N, there is a non-degenerate distribution $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ such that if we randomly pick an IID sample $(z_1, \ldots, z_n)$ according to $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ and produce a sample $S_{n-1} = (x_1, y_1), \ldots, (x_n, y_n)$ where $y_i = 1$ iff $x_i \in V_1^\omega$ then:*

$$\mathop{\mathbb{E}}_{S_{n-1} \sim P(\cdot|V_1^\omega \cup V_{-1}^\omega)} \mathop{\mathbb{E}}_{(\boldsymbol{x},y) \sim P(\cdot|V_1^\omega \cup V_{-1}^\omega)} [\chi_A(S_{n-1}, (\boldsymbol{x}, y))] > \frac{1}{4}\left(1 - \frac{n^2 - n}{N}\right)$$

The statement will follow immediately from Lemma 22 and the following claim:

Lemma 24 *Let P be a distribution over labeled samples (x, y) with the following property:*

1. *P is supported on the union of N disjoint sets $B_1, \ldots, B_N$ (i.e. $B_i \cap B_j = \emptyset$ if $i \neq j$)*
2. *For each set B_i we have that $P(y = 1|\{x_i \in B_i\}) = 1$ or $P(y = -1|\{x_i \in B_i\}) = 1$.*
3. *We have $P(\{x \in B_1\}) = P(\{x \in B_2\}) = \ldots = P(\{x \in B_N\}) = \frac{1}{N}$.*

Let $\boldsymbol{z}^{(N)}$ be a random variable such that $\boldsymbol{z}_i \in B_i$ is distributed according to $P(\cdot|B_i)$. Let $\boldsymbol{t}$ be a random variable of a subset of n elements out of N without repetition (drawn uniformly). Let $S_{n-1}(\boldsymbol{z}^{(N)}, \boldsymbol{t}) = ((\boldsymbol{z}_{t_2}, y_2), \ldots, (\boldsymbol{z}_{t_n}, y_n))$ where $y_i = 1$ iff $P(y = 1|B_{t_i}) = 1$. If

$$\mathop{\mathbb{E}}_{\boldsymbol{z}^{(N)}} \left[\ell_n(\boldsymbol{z}^{(N)}, \boldsymbol{y})\right] > a$$

then when we pick n IID elements according to P we have:

$$\mathop{\mathbb{E}}_{S_{n-1} \sim P} \mathop{\mathbb{E}}_{(\boldsymbol{x},y) \sim P} [\chi_A(S_{n-1}, (\boldsymbol{x}, y))] > \left(1 - \frac{n^2 - n}{2N}\right) a$$

Proof Consider the event E of picking a IID sample points $((\boldsymbol{z}_1, y_1), \ldots, (\boldsymbol{z}_n, y_n))$ according to P and having each $\boldsymbol{z}_i$ belong to a set B_{k_i} such that $k_i \neq k_j$ for every $i \neq j$. It is easy to see that choosing a sample set according to the distribution $P(\cdot|E)$ is exactly the random process described in the statement. Now for any positive random variable we have $E_{x \sim P}(X) \geq P(E) \cdot E_{x \sim P(\cdot|E)}(X)$ so it is enough to show that $1 - P(E) \leq \frac{n^2-n}{2N}$. Now since the sets have uniform probability, we can bound the probability that two points come from the same set as follows: pick n IID points out of N points, and let $X_{i,j}$ be a random variable such that $X_{i,j} = 1$ if $n_i = n_j$. The expected number of repetitions when drawing n points out of N points is $\mathbb{E}\left[\sum_{i<j} X_{i,j}\right] = \frac{n^2-n}{2N}$. By Markov's inequality:

$$P(\sum_{i<j} X_{i,j} \geq 1) \leq \frac{n^2 - n}{2N}.$$

■

Proof [proof of Theorem 23] For each A and N we pick V_1^ω, V_{-1}^ω and $\boldsymbol{y}$ as in Lemma 22 and draw points according to the distribution $P_{F_{\boldsymbol{y}}}(\cdot|\omega)$. Now we consider a distribution $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ where we randomly pick a ball B_i and then pick points $\boldsymbol{z}_i$ according to $P_{F_{\boldsymbol{y}}}(\cdot|\omega)$ conditioned on $\boldsymbol{z}_i \in B_i$.

Clearly $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ has the properties described in Lemma 24, and the random process that is induced by gives rise to $P_{F_{\boldsymbol{y}}}(\cdot|\omega)$ which is lower bounded by $\frac{1}{4}$.

Finally, recall that $P(\cdot|\omega)$ is supported on $B_1 \cap V_{\boldsymbol{y}_1}^\omega \times B_2 \cap V_{\boldsymbol{y}_2}^\omega \times \ldots \times B_N \cap V_{\boldsymbol{y}_N}^\omega$ and is comparable to the $N \cdot m - N$ Hausdorff measure. This in turn means that $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ is comparable to the $m-1$ Hausdorff measure and no manifold of strictly lower dimension will have positive measure, hence it is non degenerate.

■

B.4. Proof of Theorem 11

We are now ready to complete the proof of Theorem 11. By observing that for every ω we have that $V_1^\omega \in \mathcal{G}_N$ and that $P(\cdot|V_1^\omega \cup V_{-1}^\omega)$ is supported on two irreducible algebraic sets and is non degenerate hence , by Theorem 19 in $\mathcal{P}_2$, the statement is a direct consequence of Theorem 23.

B.5. Sample Complexity Lower Bounds

We now translate Theorem 11 to lower bounds on sample complexity:

Theorem 25 *For every constant C and a supervised algorithm A there is a distribution $P \in \mathcal{P}_2$, N and a hypothesis h_V $h \in \mathcal{G}_N$ such that for $\epsilon < \frac{1}{8}$ and $\delta < \frac{1}{9}$*

$$m(A, \mathcal{G}_N, P_{h_V}, \epsilon, \delta,) > C$$

Proof Pick N such that $(1 - \frac{1}{\sqrt{N}})\frac{1}{4} > \frac{2}{9}$ and $\sqrt[4]{N} > C$. For every learner A we can construct a distribution $P \in \mathcal{P}_2$ and a hypothesis $V \in \mathcal{G}_N$ such that, with $n < \sqrt[4]{N}$ examples we have

$$\mathbb{E}\left[\chi_A(S_n, (\boldsymbol{x}, y)\right] > (1 - \frac{1}{\sqrt{N}})\frac{1}{4} > \frac{2}{9}$$

But if with probability $\frac{8}{9}$ we have error smaller than $\frac{1}{8}$ the expected error will be smaller than

$$\mathbb{E}\left[\chi_A\right] < \frac{1}{9} + \frac{1}{9} < \frac{2}{9}$$

■

B.6. Final comments and proof of main result

Finally, we relate our theorems to the main results. Corrolary 25 demonstrates that there is a distribution in $P \in \mathcal{P}_2$ where the sample complexity of a supervised learner can be arbitrarily large, in particular larger than the bounds in Equation 3. As a semi-supervised learner B we choose any $\mathcal{P}_2$-ERM algorithm. B is then guaranteed by Theorem 9 and the fact that $P \in \mathcal{P}_2$ to achieve the bounds in Equation 2. This proves Theorem 5. We mention that the lower bound holds in particular for the agnostic case, and upper bounds for the agnostic case are derived in a similar manner as in the realizable case, thus we can also demonstrate that a $\mathcal{P}_2$-ERM algorithm benefits agnostically.

B.7. Leftover proofs

B.7.1. PROOF OF LEMMA 21

Let $V = \{\boldsymbol{x} : p(x_1, x_2) = 0\}$. Suppose we can write $V = V_1 \cup V_2$. Note that there are infinitely many points in V (since for every x_1 there is some x_2 such that $p(x_1, x_2) = 0$). wlog we assume there are infinitely many points in V_1. Now, suppose V_1 is the set of common roots of the polynomials $p_1(t_1, t_2), \ldots, p_k(t_1, t_2)$. For every $i \leq k$ we have that for infinitely many points:

$$p_i(x_1, f(x_1)) = 0,$$

where $f(T) = \frac{1}{\beta}\sum_{i=1}^{k} \alpha_i t^i$. Now $p_i(t, f(t))$ is a univariate polynomial with infinitely many zeroes, hence it must be identically zero and we have that for every (x_1, x_2) such that $f(x_1) = x_2$ we have $p_i(x_1, x_2) = p_i(x_1, f(x_1)) = 0$. In other words, the set V is contained in the common roots of $p_i(t_1, t_2)$ and $V \subseteq V_1$ hence $V = V_1$.

B.7.2. PROOF OF EQUATION 12

We now turn to computing JF. recall that if $F(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_N) = F_1 \ldots, F_N$ then we have by definition:

$$\boldsymbol{z}_{i,2} = \sum_{k=1}^{N} F_k(\boldsymbol{z}^{(n)})(\boldsymbol{z}_{i,1})^k$$

For every i and j we have:

$$\delta_{i,j} = \sum_{k=1}^{N} \frac{\partial F_k}{\partial \boldsymbol{z}_{j,2}}(\boldsymbol{z}_{i,1})^k \tag{15}$$

and by applying the chain rule we also have

$$\sum_{k=1}^{N} \frac{\partial F_k(\boldsymbol{z}^{(n)})}{\partial \boldsymbol{z}_{j,1}}(\boldsymbol{z}_{i,1})^k = -\delta_{i,j}\sum_{k=1}^{N} kF_k(\boldsymbol{z}^{(n)})\partial \boldsymbol{z}_{j,1}(\boldsymbol{z}_{i,1})^{k-1} \tag{16}$$

where $\delta_{i,j} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$. Taken together we have that if DF is a the Jacobian matrix of F, hence its columns are $\nabla F_1 \ldots \nabla F_n$ and $M(\boldsymbol{z}^{(n)})$ is a matrix such that $(M(\boldsymbol{z}^{(n)})_{i,k} = (\boldsymbol{z}_{i,1})^k$ then

$$MDF = W$$

where W is a matrix whose i-th column is:

$$w_i = e_{2i} - a_i e_{(2i-1)+1}$$

where $a_i = \sum_{k=1}^{N} kF_k(\boldsymbol{z}^{(n)})(\boldsymbol{z}_{i,1})^{k-1}$ and $\{e_k\}_{k=1}^{N}$ is the standard basis of $\mathbb{R}^N$. Next,

$$\sqrt{\det(MDFDF^\top M^\top)} = \sqrt{\det(M)\det(DFDF^\top)\det(M)} = \det(M)JF.$$

On the other hand:

$$\prod_{k=1}^{N} \sqrt{\left(\sum_{k=1}^{N} k F_k(\boldsymbol{z}^{(n)}) {\boldsymbol{z}_{k,2}}^{k-1}\right)^2 + 1.} = \sqrt{\det(W)} = \sqrt{\det(MDFDF^\top M^\top)}$$