IMPROVABILITY THROUGH SEMI-SUPERVISED LEARNING: A SURVEY OF THEORETICAL RESULTS

A PREPRINT

Alexander Mey
Delft University of Technology, The Netherlands
a.mey@tudelft.nl

Marco Loog
Delft University of Technology, The Netherlands
University of Copenhagen, Denmark
m.loog@tudelft.nl

October 31, 2019

ABSTRACT

Semi-supervised learning is a setting where one has labeled and unlabeled data available. In this survey we explore different types of theoretical results when one uses unlabeled data in classification and regression tasks. Most methods that use unlabeled data rely on certain assumptions about the data distribution. When those assumptions are not met in reality, including unlabeled data may actually decrease performance. Studying such methods, it therefore is particularly important to have an understanding of the underlying theory. In this review we gather results about the possible gains one can achieve when using semi-supervised learning as well as results about the limits of such methods. More precisely, this review collects the answers to the following questions: What are, in terms of improving supervised methods, the limits of semi-supervised learning? What are the assumptions of different methods? What can we achieve if the assumptions are true? Finally, we also discuss the biggest bottleneck of semi-supervised learning, namely the assumptions they make.

1 Introduction and Scope

In various applications gathering unlabeled data is easier, faster and/or cheaper than gathering labeled data. The goal of semi-supervised learning (SSL)[1] is to combine unlabeled and labeled data to design classification or regression rules that outperform schemes that are only based on labeled data. SSL does come, however, with an inherent risk. It is well-known that including unlabeled data can degrade the performance [Ben-David et al., 2008, Cozman and Cohen, 2006]. Studying and understanding SSL from a theoretical point of view allows us to exactly formulate the assumptions we need and the improvements we can expect, as well as the limitations of said methods. With this one can formulate recommendations for using SSL with the aim of avoiding a decrease in performance as good as possible. In this review, we collect and present theoretical results concerning SSL, study the relevant papers in detail, present their main result and point out connections to other works.

This review targets two groups of audience. The first group we target are interested practitioners and researchers working on experimental SSL. While they may not be interested in all the details we present, we believe that the introduction in each of our sections gives a good high level understanding of the types of theoretical results in SSL and the main insights they provoke. The second target audience is everyone working on the theoretical side of SSL. We hope that, especially researchers starting in this field, can find inspiration and connections to their own work in our overview. We mostly present results that describe the performance of semi-supervised learners, often, but not exclusively, in the language of the PAC-learning framework.[2] We interpret the results, draw connections between them and point out what

[1] We overload the abbreviation of SSL to stand either for *semi-supervised learning* or *semi-supervised learner*.

[2] PAC-learning stands for *Probabilistically Approximately Correct*-learning. In this framework one can study how far a trained classifier is off of the best classifier from a given class, given a certain amount of labeled data. The rate at which we approach the best classifier is called learning rate. Nice introductions to this framework can be found in Shalev-Shwartz and Ben-David [2014] and Mohri et al. [2012]. We also refer to Definition 1, where we introduce the notion of sample complexity. PAC-learnable means that the sample complexity is always finite.

one has to assume for them to be valid. Next to theoretical guarantees of some specific SSL we also present results on the limits of SSL.

1.1 Outline

In the next section we introduce the formal learning framework which is also assumed for the majority of the work we present. In Section 3 we present results on the limits of SSL, which often arise due to specific assumptions on the model or the data generation process. Opposing to the settings where the improvements of SSL are provably limited, we present in the same section three settings where the improvements of SSL are *unlimited*. With unlimited we mean here that a SSL can PAC-learn the problem, while no supervised learner (SL) can. In Section 4 we investigate methods that try to exploit unlabeled data, without having further assumptions on the data distribution. In Section 5 we present semi-supervised learners that make *weak* assumptions on the data distribution. Those assumptions are weak in the sense that the resulting learner cannot get a learning rate faster than the standard learning rate of $\frac{1}{\sqrt{n}}$,[3] where n is the number of labeled samples. The improvements are instead given by a constant. In Section 6 we present learners that use *strong* assumptions under which one can converge exponentially fast to the best classifier in a given class, i.e. the learning rate is in order of e^{-n}. In Section 7 we present results in the transductive setting, a setting where one is only interested in the labels of the unlabeled data. In the same section we also present a line of research that tries to construct semi-supervised learners that are never worse than their supervised counterparts. In Section 8 we discuss the overall results and point out what the current challenges in the field are. In Section 8.4 we furthermore explain in more detail what is formally meant by using assumptions in SSL and the problems that occur with that.

2 Preliminaries

Unless further specified all results are presented in the standard statistical learning framework. This means that we are given a feature space $\mathcal{X}$ and a label space $\mathcal{Y}$ together with an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$. Overloading the notation we write $P(X)$ and $P(Y)$ for the marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$ and similar for conditional distributions. We observe a labeled n-sample $S_n = ((x_1, y_1), ..., (x_n, y_n))$ and an unlabeled m-sample $U_m = (x_{n+1}, ..., x_{n+m})$, where each (x_i, y_i) for $1 \leq i \leq n$ and each x_j for $n+1 \leq j \leq n+m$ is identically and independently distributed according to P. One then choses a hypothesis class H, where each $h \in H$ is a mapping $h : \mathcal{X} \to \mathcal{Y}$, and a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. Unless specified otherwise we assume for classification that $\mathcal{Y} = \{-1, +1\}$ and the loss is the 0-1 loss, $l(y, \hat{y}) = I_{\{y \neq \hat{y}\}}$. In the regression task we assume that $\mathcal{Y} = \mathbb{R}$ and $l(y, \hat{y}) = (y - \hat{y})^2$. Based on the n labeled and m unlabeled samples we then try to find a $h \in H$ such that the risk $R(h) := \mathbb{E}_{X,Y}\left[l(h(X), Y)\right]$ is small. Finally, whenever we have any quantity A that depends on the distribution P, we write $\hat{A}$ for a empirically estimated version of A. For example, given a labeled sample S_n we write $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$ for the empirical risk of $h \in H$ measured on S_n. If not clear from context we will clarify on which sample we measure. In Table 1 on page 24 we present a complete list of the notation we use.

3 Possibility & Impossibility of Semi-Supervised Learning

In SSL we want to use information about the distribution on $\mathcal{X}$ to improve learning, but it is not necessarily clear that this information can be useful at all. Some authors formalize this idea and then present situations where unlabeled data can help or where it cannot. This section follows the same division. In Subsection 3.1 we present different settings where authors could show that unlabeled data cannot help, while In Subsection 3.2 we present three specific settings where unlabeled data can give unlimited improvements. By unlimited we mean that no supervised learner can PAC learn in those settings while a semi-supervised learner can.

The negative results often assert an independence between the posterior probability $P(Y \mid X)$ and the marginal distribution $P(X)$. This does, however, not directly mean that unlabeled data is useless, as we are usually not only interested in $P(Y \mid X)$ but on the complete risk of a classifier h, $\mathbb{E}_{X,Y}\left[l(h(X), Y)\right]$, which *does* depend on $P(X)$ [Peters et al., 2017, 5.1.2]. In Section 4.1 and 4.2, for example,we present work that show risk improvements even when $P(Y \mid X)$ and $P(X)$ are independent.

3.1 Impossibility Results

3.1.1 Impossibility Because of the Data Generation Process

Seeger [2000] looks at a simple data generation model and investigates how prior information about the data distribution changes our posterior belief about the model if the prior information is included in a Bayesian fashion. To use the Bayesian approach, the data is assumed to be generated in the following manner. We assume now that the distribution P comes from a model class with parameters μ and θ. First values $\mu \sim P_\mu$ and $\theta \sim P_\theta$ are sampled independently and then the data is generated by gathering samples $x \sim P(X \mid \mu)$ with corresponding labels $y \sim P(Y \mid X, \theta)$ as shown in Figure 1. The goal in this setting is to infer θ from a finite labeled sample $S_n = (x_i, y_i)_{1 \leq i \leq n}$. Using a Bayesian approach it can be easily shown that $P(\theta \mid S_n)$ is independent of any finite unlabeled sample and μ itself. In other words: Unlabeled information does not change the posterior belief about θ given the labeled data S_n. A possible solution presented is to assume a dependency between μ and θ, so drawing an additional arrow between μ and θ in Figure 1.

3.1.2 Impossibility Because of The Model Assumptions

Hansen [2009] investigates when unlabeled data should change our posterior belief about a model. In comparison to Seeger [2000] no data generation assumptions are made, but rather assumptions about the model we use. He looks at solutions derived from the expected squared loss between this given model and the true desired label output. Splitting the joint distribution $P(X, Y \mid \theta)$ of our model as $P(X, Y \mid \theta) = P(Y \mid X, \theta_1, \theta_2)P(X \mid \theta_2, \theta_3)$ he concludes that unlabeled data can be discarded if θ_2, the shared parameter between the label and marginal distribution, is empty.

Earlier work by Zhang and Oles [2000] distinguishes the same type of models, but the impossibility is about the asymptotic efficiency of semi-supervised classifiers. The paper as well considers two types of joint probability models:

1. Parametric: $P(X, Y \mid \alpha) = P(X \mid \alpha)P(Y \mid X, \alpha)$
2. Semi-Parametric: $P(X, Y \mid \alpha) = P(X)P(Y \mid X, \alpha)$

One can show that the Fisher information $I(\hat{\alpha})_{\text{unlabeled + labeled}}$ of an MLE estimator $\hat{\alpha}$ that takes labeled an unlabeled data into account can be decomposed as $I(\hat{\alpha})_{\text{unlabeled +labeled}} = I(\hat{\alpha})_{\text{unlabeled}} + I(\hat{\alpha})_{\text{labeled}}$. So, as long as unlabeled data is available, the Fisher information of the semi-supervised learner is bigger compared to the supervised learner, which is shown to have a Fisher information given by $I(\hat{\alpha})_{\text{labeled}}$. It follows that the SSL is asymptotically more efficient, although not necessarily strictly. In the parametric case we observe that $I(\hat{\alpha})_{\text{unlabeled}} = 0$ and the semi-supervised and supervised estimator have the same asymptotic behavior. In Section 4.1 we will present a method that allows for asymptotic efficiency of a SSL even when using a discriminative model $P(Y \mid X, \alpha)$.

3.1.3 Impossibility Because of The Causal Direction

Schölkopf et al. [2012, Sections 2 and 3] analyze a functional causal model shown as in Figure 2. They analyze different learning scenarios under the assumption that the label is the cause C and the feature is the effect E and vise versa. This model introduces an asymmetry in cause and effect, since it leads to the fact that $P(C)$ and $P(E \mid C)$ are independent, while $P(E)$ and $P(C \mid E)$ are not independent. Assuming now that X is the cause of the label Y, we find that the prediction $P(Y \mid X)$ is independent of newly gained information about $P(X)$. The situation changes though if we assume that the label Y was caused by X. One problem with this is, that we do not necessarily know if the feature is a cause or an effect. But for example in medical settings this might not be too difficult, as we can identify causal features as those that do actually cause an illness, while effect features are the symptoms of an illness. Kügelgen et al. [2019] use this knowledge and derive a SSL method which only takes the unlabeled data of effect features into account.

$$\begin{tikzcd}
\mu \arrow[d] & \theta \arrow[d] \\
X \arrow[r] & Y
\end{tikzcd}$$

Figure 1: The data generation process used in the analysis of Seeger.

[3] The learning rate is the rate in which we converge to the best classifier from a given class in number of the labeled samples. That the standard rate is in order of $\frac{1}{\sqrt{n}}$ follows from classic results as shown for example by Vapnik [1998].

$$\begin{tikzcd}
N_c \arrow[d] & N_e \arrow[d] \\
C \arrow[r, "\varrho"] & E
\end{tikzcd}$$

Figure 2: Simple functional causal model used by Schölkopf et al. [2012]. The effect E is caused by C given a deterministic mapping ϱ. Both E and C are influenced by a noise variables N_E and N_C.

3.1.4 Impossibility to Always Outperform a Supervised Learner

Inspired by a successful minimax approach for a *generative* linear discriminant model of Loog (see Section 7.2.2), Krijthe and Loog [2018] investigate a similar approach to find semi-supervised solutions for *discriminative* models that are never worse than their supervised counterparts. They use a setting where the discriminative models are derived with a monotonously decreasing loss function. The setting is also transductive, so where one is only interested in the performance of our model on the unlabeled data U_m, see also Section 7. They essentially show that, under some mild conditions, there is always a labeling of the unseen data U_m such that a semi-supervised learner will perform worse on U_m than the supervised solution. In this sense it is impossible to guarantee that the semi-supervised solution will always outperform the supervised solution.

3.1.5 Impossibility If We Only Know the Manifold

Lafferty and Wasserman [2007, Section 3] show that knowledge of the manifold alone, without additional assumption, is not sufficient to outperform a purely supervised learner. They work in a regression setting and extend work of Bickel and Li [2007] to show that there is a supervised learner that can adapt to the dimension of the manifold and thus can achieve minimax rates equivalent to a learner that directly works on the lower dimensional manifold.

We note that Lafferty and Wasserman [2007] also show that we can essentially achieve faster rates if we also assume a semi-supervised smoothness assumption. We do not cover more details at this point, but offer a qualitatively very similar analysis in Section 6.4.

3.1.6 Impossibility If We Don't Have Additional Assumptions

Ben-David et al. [2008] started a series of investigations by conjecturing that SSL is, in some sense, generally not possible without any assumptions. In particular we assume that a given domain distribution does not restrict the possible labeling functions, similarly to the data generation process in Figure 1. They hypothesize that a semi-supervised learner can't have essentially better sample complexity bounds (see Definitions 1 and 2) than a SL, without any additional assumptions at least. This is different from the previous sections, as there are no further restrictions on the model or the data generation process.

In the following two sections we want to illustrate the precise idea of those conjectures, why they do not hold generally and in which scenarios they are true.

We start with the contributions of Ben-David et al. [2008]. They hypothesize that the worst-case sample complexity for any semi-supervised learner improves over a supervised learner at most by a constant which only depends on the hypothesis class. The first conjecture states that for the realizable case.

Conjecture 1 (Conjecture 4). [4] *For any hypothesis class H, there exists a constant $c(H)$ such that for any domain distribution D on $\mathcal{X}$*

$$\sup_{h \in H} m(H, D_h, \epsilon, \delta) \leq \sup_{h \in H} c(H) m^{\text{SSL}}(H, D_h, \epsilon, \delta), \tag{1}$$

for ϵ and δ small enough, where D_h is the distribution on $\mathcal{X} \times \mathcal{Y}$ with marginal distribution D and conditional distribution $D_h(Y = h(x) \mid X = x) = 1$.

The second conjecture states the same for the agnostic case, so where we replace D_h for any arbitrary distribution P.

Conjecture 2 (Conjecture 5). *For any hypothesis class H, there exists a constant $c(H)$ such that for any domain distribution D*

$$\sup_{P \in \text{ext}(D)} m(H, P, \epsilon, \delta) \leq \sup_{P \in \text{ext}(D)} c(H) m^{\text{SSL}}(H, P, \epsilon, \delta), \tag{2}$$

for ϵ and δ small enough and where $\text{ext}(D)$ *is the set of all distributions P on $\mathcal{X} \times \mathcal{Y}$ such that the marginal distribution fulfills $P(X) = D$.*

[4] In brackets we note under which name the statement can be found in the original paper.

In other words: The paper conjectures that if we are given a fixed domain distribution, one can always find a labeling function on it such that for this labeling function the sample complexity gap between SL and SSL can only be a constant. The paper proofs these conjectures for smooth distributions on the real line and threshold functions in the realizable case and for threshold functions and unions of intervals in the agnostic case. The sample complexity comparison is by construction a worst case analysis, in cases where the target hypothesis behaves benign we might still get non-constant improvements. We explore those cases in Section 6. On another note, one can also ask the question how good a constant improvement by itself can already be. We will elaborate on this in the discussion.

The Conjectures 1 and 2 are essentially true in the realizable case when the hypothesis class has finite VC-dimension. Darnstädt et al. [2013] showed that Conjecture 1, the realizable case, is true with a small alteration: the supervised learner is allowed to be twice as inaccurate and for the finite VC-dimension case we get an additional term of $\log(\frac{1}{\epsilon})$. Mey et al. [2019] take this idea, in a certain way, a step further, and show that a manifold regularization scheme obeys the limits stated by the conjecture, even though in this case the domain distribution carries information about the labeling function. Darnstädt et al. [2013] prove the following version of Conjecture 1.

Theorem 1 (Theorem 1)**.** *Let H be a hypothesis class such that it contains the constant zero and constant one function. Then for every domain distribution D and every $h \in H$,*

1. *If H is finite then*
$$m(H, D_h, 2\epsilon, \delta) \leq O(\ln |H|) m^{\mathrm{SSL}}(H, D_h, \epsilon, \delta). \tag{3}$$

2. *If H has finite VC-dimension then*
$$m(H, D_h, 2\epsilon, \delta) \leq O(\mathrm{VC}(H)) \log(\frac{1}{\epsilon}) m^{\mathrm{SSL}}(H, D_h, \epsilon, \delta). \tag{4}$$

First note that this statement holds for all D_h, so in particular if we take the supremum over all $h \in H$ as in Conjecture 1. Golovnev et al. [2019] show that if the hypothesis class H is given by the projections over $\{0,1\}^d$, there is a set of domain distributions such that any supervised algorithm needs $\Omega(\mathrm{VC}(H))$ as many samples as the semi-supervised counterpart, which has knowledge of the full domain distribution. So in particular Inequality (4) is tight up to logarithmic factors. This actually shows that the constant improvement can be arbitrarily good, as we can increase the VC-dimension by increasing the dimension Golovnev et al. [2019, Proposition 4]. The agnostic version of Theorem 1 is an open problem.

In the case of a hypothesis class with infinite VC-dimension, however, the conjecture ceases to hold, also for the slightly altered formulations. This is essentially the case because we can start with a class that has infinite VC-dimension, and thus cannot be learned by a supervised learner. A semi-supervised learner, however, can restrict this class in a way such that it has finite VC-dimension. This will become clearer in the next section where we collect three different setups in which a semi-supervised learner can PAC-learn, while a supervised learner cannot.[5]

3.1.7 Impossibility If We Don't Restrict The Possible Labeling Functions

Golovnev et al. [2019] show that if the domain $\mathcal{X}$ is finite and we allow all deterministic labeling functions on it, no semi-supervised learner can improve in the realizable PAC-learning framework even by a constant over a consistent supervised learner. Consistent means here that the learner achieves 0 training error. The supervised learner is, however, to be allowed twice as inaccurate and twice as unsure.

Theorem 2 (Theorem 8)**.** *Let $\mathcal{X}$ be a finite domain, and let $H_{\text{all}} = \{0,1\}^{\mathcal{X}}$ be the set of all deterministic binary labeling functions on $\mathcal{X}$. Let A be any consistent supervised learner, P a distribution over $\mathcal{X}$ and $\epsilon, \delta \in (0,1)$. Then*
$$m(A, H_{\text{all}}, P, 2\epsilon, 2\delta) \leq m^{\mathrm{SSL}}(H_{\text{all}}, P, \epsilon, \delta). \tag{5}$$

While the more general Theorem 1 states that a semi-supervised can still be better by a constant depending on the hypothesis class, we find that in the previous setting one even loses this advantage.

A similar result can be found for the agnostic case. Theorem 2 of Göpfert et al. [2019] essentially states that Conjecture 2 (the agnostic case), is true for the finite VC-dimension case, if there are no restrictions on the labeling function. The difference is that they consider in an in-expectation and not a high probability framework and there is a condition on the domain distribution D, while Conjecture 2 is formulated to hold for *all* distributions D. This condition is, however, very mild, the essential assumption of the theorem is that there are no restrictions on the labeling function.

The intuition for both of the previous results is the same: If we allow all labeling functions, there is no label information about the support of $\mathcal{X}$ that we did not observe yet. Finding the labels for this part is equally slow for supervised and

[5] In this context PAC-learnability means that $m(H, \epsilon, \delta)$ is finite for all $\epsilon, \delta > 0$.

semi-supervised learners. In the next section we present hypothesis classes on which semi-supervised learners can be effective. Following the previous result, it is not surprising that those classes are carefully chosen.

3.2 Proofs about the Possibility of Semi-Supervised Learning

We consider three specific settings in which it can be shown that a SSL can learn, while a SL cannot. We first present the work of Darnstädt et al. [2013] and Globerson et al. [2017], these aim to answer Conjectures 1 and 2 covered in the previous subsection. They show that there is a hypothesis class H^* and a collection of domain distributions $\mathcal{D}^*$ such that no supervised learner can learn H^* under the distributions of $\mathcal{D}^*$. Given, however, any $P \in \mathcal{D}^*$, a semi-supervised learner that has access to a finite, but depending on P arbitrarily large, amount of unlabeled data can learn H^* with the same rate of convergence. Next we present the work of Niyogi [2013] as it gives the best example to illustrate how a shift from not learnable to learnable is possible when going from SL and SSL.

3.2.1 Proving the Realizable Case with a Discrete Set

Darnstädt et al. [2013] give the first example that shows that Conjecture 1 does not generally hold. This is captured in the following theorem, and the other results of this section will be very similar.

Theorem 3 (Theorem 2)**.** *There exists a hypothesis class H^* and a family of domain distributions $\mathcal{D}^*$ such that*

1. *For every $D \in \mathcal{D}^*$,*
$$m^{\text{SSL}}(H^*, D, \epsilon, \delta) \leq O(\frac{1}{\epsilon^2} + \frac{1}{\epsilon}\log(\frac{1}{\epsilon})).$$
2. *For all $\epsilon < \frac{1}{2}$ and $\delta < 1$,*
$$m(H^*, \epsilon, \delta) = \sup_{D \in \mathcal{D}^*} m(H^*, D, \epsilon, \delta) = \infty.$$

In order for the SSL to be able to PAC-learn for all $D \in \mathcal{D}^*$ it needs knowledge of the full distribution D. (Although for each fixed $D \in \mathcal{D}^*$ a finite amount of unlabeled data suffices). Since the supervised learner can only collect labeled samples it will never be able to achieve this knowledge with a finite number of samples, and thus has an infinite sample complexity. The construction of H^* and $\mathcal{D}^*$ can be considered rather artificial. We discuss papers that show similar behavior with a hypothesis class which is loosely based on the manifold assumption in the next two subsections. We nevertheless want to give the intuition for the given example, as it, as well as the other examples, use the same trick.

Darnstädt et al. [2013] set the example up as follows. The domain $\mathcal{X}$ consists of all sequences $x = (x_1, x_2, ..., x_l)$ of arbitrary finite length and $x_i \in \{0, 1\}$. The distributions $D \in \mathcal{D}^*$ on $\mathcal{X}$ are such that there is a sequence $D(x_{\sigma(1)} = 1) > D(x_{\sigma(2)} = 1) > ...$, which drops sufficiently quick[6], where σ is a random permutation on the length of x. The hypothesis class H^* contains all hypotheses h_i with $h_i(x) = x_i$ and the constant 0 hypothesis. Note that although the class has infinite VC-dimension it still takes some effort to show that no supervised learner can learn it w.r.t to all distributions in $\mathcal{D}^*$. This is because the VC-dimension might not be infinite over $\mathcal{D}^*$. We want to sketch how the SSL can learn it. After fixing a $D \in \mathcal{D}^*$ and $\epsilon, \delta > 0$ we draw enough unlabeled samples to identify all positions $i \in \mathbb{N}$ such that x_i is with a high probability 0. For all those indices i we can remove h_i from H^* as the constant 0 hypothesis will be good enough for predicting accurately. They then show that the remaining hypotheses in H^* can be learned from finitely many samples. Note that it is important that the admissible domain distributions are restricted. If D^* would also include distributions that essentially put equal weight on all positions i, unlabeled data could not help to restrict H^*. In short: this example, and also the following, are essentially set up such that H and D have a certain link, and in those cases knowledge about D can actually give knowledge about H. Note, however, that the knowledge about D did not restrict the set of possible labeling functions from H. It was rather that D helped to identify which hypotheses we can safely ignore.

3.2.2 Proving the Agnostic Case using Algebraic Varieties

Globerson et al. [2017] provide a different example using a hypothesis class which loosely follows the manifold assumption. Using the same example one can also show that Conjecture 2, so the impossibility conjecture for the agnostic case, is not true in general.

The theorem is very similar to Darnstädt et al. [2013], the difference is in the construction of the hypothesis set and the set of distributions.

[6] Note that with $x_{\sigma(i)} = 1$ we mean the subset $V \subset \mathcal{X}$ with $V := \{x = (x_1, x_2, ..., x_l) \in \mathcal{X} \mid x_{\sigma(i)} = 1\}$.

Theorem 4 (Theorem 5)**.** *There exists a hypothesis class H_{alg} and a set of distributions $\mathcal{D}_{\text{alg}}$ such that.*

1. *For every $D \in \mathcal{D}_{\text{alg}}$,*
$$m^{\text{SSL}}(H_{\text{alg}}, D, \epsilon, \delta) < \frac{2}{\epsilon} \log \frac{2}{\delta}. \tag{6}$$
2. *The supervised sample complexity is infinite,*
$$\sup_{D \in \mathcal{D}_{\text{alg}}} m(H_{\text{alg}}, D, \epsilon, \delta) = \infty. \tag{7}$$

The hypothesis class H_{alg} consists of all hypotheses that have class label 1 on an algebraic set, so essentially a type of manifold, and 0 outside of that algebraic set. This is still a very expressive set with infinite VC dimension. But if we restrict the set of admissible domain distributions $\mathcal{D}_{\text{alg}}$ also to be (a certain type of) algebraic sets, a semi-supervised learner with knowledge of $D \in \mathcal{D}_{\text{alg}}$ can learn efficiently: we can think of $\mathcal{D}_{\text{alg}}$ as the set of distributions that have support on a finite combination of distinguishable algebraic sets $V_1, ..., V_k$. Once we know that the distribution has support on $V_1, ..., V_k$, we only have to figure out which of those algebraic sets have label 1 and which have label 0. A SSL can thus reduce the class H_{alg} by only considering the hypotheses that have class label 1 on combinations from $V_1, ..., V_k$. Since the set of all possible combinations is finite, a SSL can learn them with a sample complexity bounded by Inequality (6). Note that although the true labeling function does not have to be part of this restricted set, one can show that it is anyway always optimal to predict with a hypothesis from it. The argument for that is similar to the explanation of the agnostic case below.

The paper also discusses that this argumentation can be extended to the agnostic case, so when the true target function is not in H_{alg}. This extension might appear problematic at first, because the semi-supervised algorithm restricts the hypothesis set H_{alg}, and to guarantee PAC-learnability we need to know that the best predictor from the H_{alg} is still in this restricted set. But this is indeed the case, because the set of domain distributions $\mathcal{D}_{\text{alg}}$ was exactly created for that to hold. To show that, assume that the distribution is supported on an irreducible algebraic set V_0. Our SSL can now chose to label it completely 1 or 0, while both options might lead to non-zero error. But labeling it completely as either 1 or 0 is already ideal, as using any other algebraic set $V_1 \in H_{\text{alg}}$ will lead to one of those two labelings. This is because, by construction, V_1 is either equal to V_0 (which leads to label everything as 1) or has an intersection of zero mass (which leads to labeling almost everything as 0).

This seems to contradict the findings in 3.1.5, as Lafferty and Wasserman [2007] show that a supervised learner can also adapt to the underlying manifold. This discrepancy is not easy not analyze as Lafferty and Wasserman [2007] work in the regression setting, while Globerson et al. [2017] analyse classification. The intuition, however, is that Globerson et al. [2017] present the supervised learner with an impossible, meaning not PAC-learnable, task. Lafferty and Wasserman [2007] on the other hand restrict the target functions to be smooth, and thus the supervised learner is presented with a sufficiently easy problem. Mey et al. [2019] show indeed that if the supervised learner is presented with a learnable classification task, manifold regularization can improve the sample complexity only by a constant.

3.2.3 Using the Manifold Assumption to Make A Class Learnable

Niyogi [2013] provides another setup in which a semi-supervised learner can effectively learn while a supervised learner cannot. The motivation, however, was independent of Ben-David et al. [2008] and was meant as a general theoretical analysis of the manifold learning framework as introduced in Belkin et al. [2006]. Also, their results are in-expectation, while the previous papers give PAC bounds, which means that they hold with high probability. Although the paper presents the results in an in-expectation framework we slightly alter the setup and present it in the PAC learning framework. We believe this is sufficient to understand the ideas and allows us to draw better connections to the previous papers. Although this work is based on the manifold assumption, so a given domain distribution does limit the possible labeling functions, we believe that it is the most intuitive setting to understand why a supervised learner cannot learn, while a semi-supervised learner can.

The example is built as follows. First it is assumed that the admissible domain distributions are given by the class $\mathcal{P}_c$ which have support on embeddings of a circle in the Euclidean plane, see also Figure 3. The hypothesis class H_c consists of all possible binary labelings of half circles, while everything outside the circle is labeled as 1,[7] see also Figure 3. The SSL that knows the specific embedding of the circle, only needs to find two thresholds on the given circle, a class with VC-dimension of 2, so the SSL can learn efficiently. In Figure 4 we schematically show why H_c has an infinite VC dimension and thus cannot be learned by any supervised learner.

[7] The labeling outside of the circle is a formality to ensure that the supervised learner makes predictions for the whole circle, as the learner does not a priori know in which part of the space the circle is embedded.

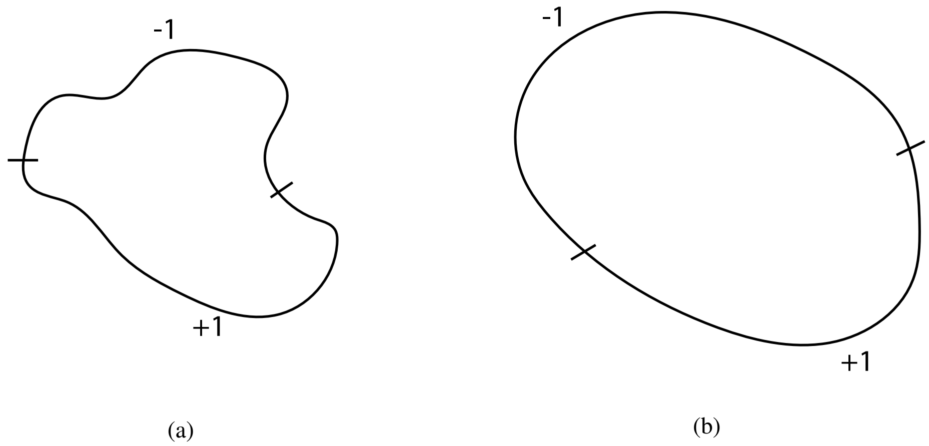
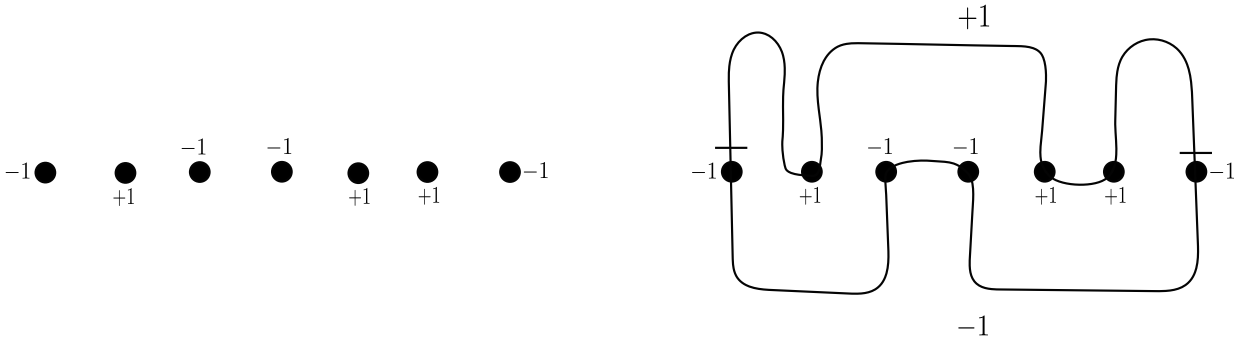


Figure 3: The shapes shown in (a) and (b) are two different embeddings of a circle in the Euclidean plane. One half of the circle is labeled as 1, while the other half is labeled as -1, while we assume that everything outside the circle is labeled as 1.



(a) Assume we are given 7 points that are labeled as depicted above.

(b) The circle above labels the points correctly. The upper half assigns points the label -1, while the lower half labels points as $+1$.

Figure 4: A schematic proof why the hypothesis set H_c has an infinite VC dimension. Given the points in (a) we can label them correctly with the circle given in (b).

4 Learning Without Assumptions

As argued in the previous section it can be difficult to use unlabeled data without any additional assumptions, and in some situations one can show that unlabeled data cannot help at all. As already mentioned in the introduction of Section 3, this impossibility stems sometimes from the fact that we only consider improvements of the estimate of the conditional probability $P(Y \mid X)$. The work we present in this section looks at the complete risk $\mathbb{E}_{X,Y}\left[l(h(X), Y)\right]$, a quantity which is always influenced by the marginal distribution $P(X)$. Furthermore no additional assumptions about the distribution P are made, and the theoretical guarantees are accordingly weak. We first present the work of Sokolovska et al. [2008] who use the unlabeled data to reweigh the labeled points, and show improvements in terms of asymptotic efficiency. Interestingly, one needs that the model is misspecified to show this result. Second we present the work of Kääriäinen [2005] who uses the unlabeled data to pick the center of the version space. The best possible improvements are bounded by a factor of 2. Finally we present the work of Leskes [2005] who uses unlabeled data to combine different hypothesis spaces and shows that the learning rates depend on the highest Rademacher complexity amongst those hypothesis spaces.

4.1 Reweighing the Labeled Data By the Marginal Distribution

Sokolovska et al. [2008] proposed a semi-supervised learner that uses knowledge of the marginal distribution $P(X)$ in a re-weighing scheme. To avoid difficulties for the theoretical analysis they restrict the feature space $\mathcal{X}$ to contain only finitely many points and assume that the SSL has access to the full marginal distribution $P(X)$.[8] They consider models that directly estimate class probabilities $p(y \mid x, \theta)$, while they measure performance by the negative log-likelihood $l(x, y \mid \theta) = -\ln p(y \mid x, \theta)$. They then analyze asymptotic behavior, in particular the asymptotic variance of the model estimation. They compare two models, the classical maximum log-likelihood estimate based on the labeled data only

$$\theta^{\text{SL}} = \arg\min_{\theta \in \Theta} \sum_{(x,y) \in S_n} l(x, y \mid \theta) \tag{8}$$

and a semi-supervised learner that also takes the marginal $P(x)$ into account

$$\theta^{\text{SSL}} = \arg\min_{\theta \in \Theta} \sum_{(x,y) \in S_n} \frac{P(x)}{\sum_{z \in X_n} I_{\{x=z\}}} l(x, y \mid \theta). \tag{9}$$

Again, note that the semi-supervised learner weighs each feature with the true, instead of the empirical, distribution. Let us first state the results about θ^{SSL} and then discuss them.

Theorem 5 (Theorem 1). *Let $\theta^* \in \arg\min_{\theta \in \Theta} \mathbb{E}[l(x, y \mid \theta)]$ and define the following matrices*

$$H(\theta^*) = \mathbb{E}_X \left[\mathbb{V}_{Y|X}[\nabla_\theta l(X, Y \mid \theta) \mid X]\right] \tag{10}$$

$$I(\theta^*) = \mathbb{E}_{X,Y} \left[\nabla_\theta l(X, Y \mid \theta) \nabla_\theta^T l(X, Y \mid \theta)\right] \tag{11}$$

$$J(\theta^*) = \mathbb{E}_{X,Y} \left[\nabla_\theta^T \nabla_\theta l(X, Y \mid \theta)\right], \tag{12}$$

where $\mathbb{V}_{Y|X}$ is the variance over the conditional random variable $Y \mid X$. Then θ^{SL} and θ^{SSL} are consistent and asymptotically normal estimators of θ^* with

$$\sqrt{n}(\theta^{\text{SL}} - \theta^*) \to \mathcal{N}(0, J^{-1}(\theta^*) I(\theta^*) J^{-1}(\theta^*)) \tag{13}$$

$$\sqrt{n}(\theta^{\text{SSL}} - \theta^*) \to \mathcal{N}(0, J^{-1}(\theta^*) H(\theta^*) J^{-1}(\theta^*)) \tag{14}$$

and θ^{SSL} is asymptotically efficient, meaning that it achieves asymptotically the smallest variance of any unbiased estimator.

Asking now when θ^{SSL} asymptotically dominates θ^{SL} we get the somewhat surprising answer that we need the model to be misspecified. From a statistical point of view it is maybe not so surprising, since in the well-specified case (along with some other regularity conditions) the MLE θ^{SSL} is already asymptotically efficient itself. Specifically, we have that then $H(\theta^*) = J(\theta^*) = I(\theta^*)$, and we recover the classical result that the MLE is asymptotically normal with a variance of the inverse Fisher information matrix $I(\theta^*)$. The paper then examines, with the logistic regression model, when the difference between $I(\theta^*)$ and $H(\theta^*)$ is particularly big. It is shown that this is the case the more $P(Y \mid X)$ is bounded away from $1/2$, so in particular when the Bayes error is small. This is very similar to *Tsybakov's low noise* Tsybakov [2004], which is used in statistical learning to show fast learning rates. In Sections 6.1 and 6.2 similar assumptions are made to show that some semi-supervised learners can converge exponentially fast to the Bayes error.

4.2 Using the Unlabeled Data to Pick the Center of the Version Space

Kääriäinen [2005] introduces a method for bounding the risk by using unlabeled data to collect information about the agreement of two classifiers. A semi-supervised estimator is then derived as the hypothesis that minimizes this bound. Unfortunately the idea only works really in the realizable case. Although we do not get a new algorithm for the agnostic case, the paper still presents new bounds based on the unlabeled data.

4.2.1 Realizable Case

The idea for the realizable case is to consider the version space, so the space that contains all hypotheses that have no training error. The unlabeled data gives rise to a pseudo-metric on this space by measuring the disagreement of the hypotheses on it. We are going to pick the hypothesis that has the lowest worst-case disagreement to all other

[8] The work is continued by Kawakita and Kanamori [2013] and extended to non-discrete features spaces.

hypothesis, of which one must be the true one as we assume realizability. Let us make this more precise. Given two hypotheses $f, g \in H$ we define the disagreement pseudo-metric $d(f, g)$ as

$$d(f, g) = P(f(X) \neq g(X)). \tag{15}$$

This metric is specifically useful in the semi-supervised case since is does not depend on labels. We can approximate it with the empirical version by

$$\hat{d}(f, g) = \frac{1}{m} \sum_{i=n}^{n+m} I_{\{f(x_i)=g(x_i)\}}. \tag{16}$$

The version space is defined as $H_0 = \{h \in H \mid \hat{R}(h) = 0\}$. Let h_0 be the true hypothesis, then we know that $h_0 \in H_0$ and one can show that $R(h) = d(h, h_0)$ for all $h \in H$. This allows us to bound

$$R(h) = d(h, h_0) = \hat{d}(h, h_0) + (\hat{d} - d)(h, h_0) \leq \sup_{g \in H_0} \hat{d}(h, g) + \sup_{g, g' \in H_0} (\hat{d} - d)(g, g'). \tag{17}$$

As Inequality (17) bounds the true risk of a hypothesis h, we try to minimize this risk by choosing the hypothesis that minimizes the right-hand side of Inequality (17). More precisely, we choose the semi-supervised estimator as the *empirical center of the version space*, so we set

$$h^{\text{SSL}} = \arg \inf_{h \in H_0} \sup_{g \in H_0} \hat{d}(h, g).$$

With this we can of course only control the first term on the right-hand side of Inequality (17). We can bound the second term, however, with concentration inequalities derived from a Rademacher Complexity for the space $\mathcal{G} = \{x \mapsto I_{\{f(x)=g(x)\}} \mid f, g \in H_0\}$. It is then true that with probability at least $1 - \delta$ [Kääriäinen, 2005, Theorem 3]

$$R(h^{\text{SSL}}) \leq \inf_{h \in H_0} \sup_{g \in H_0} \hat{d}(h, g) + \text{empRad}(\mathcal{G}) + \frac{3}{\sqrt{2}} \sqrt{\frac{\ln \frac{2}{\delta}}{m}}. \tag{18}$$

Note the two terms on the right hand-side of Inequality (18) go to 0 for increasing m and note that in this case also $\hat{d}(f, g) \to d(g, g)$. So ignoring for a minute that we only have finitely many unlabeled data we can compare the SSL (16) purely supervised solutions. Note that in the realizable case a purely supervised method would also choose a hypothesis in H_0. As the supervised learner h^{SL} has no further information we can always find a target hypothesis h^* such that $R(h^{\text{SL}}) = \sup_{g \in H_0} d(h^{\text{SL}}, g) = d(h^{\text{SL}}, h^*)$. So the best bound for any supervised learner h^{SL} is given by $R(h^{SL}) \leq \sup_{g \in H_0} d(f, g)$. The SSL bound (18) on the other hand allows us to bound $R(h^{\text{SSL}}) \leq \inf_{h \in H_0} \sup_{g \in H_0} d(h, g)$, at least for m going to infinity. From a geometrical viewpoint $\sup_{g \in H_0} d(h^{\text{SL}}, g)$ is the diameter of H_0, while, $\inf_{h \in H_0} \sup_{g \in H_0} d(h, g)$ is the radius. As the difference between the radius and the diameter, with respect to d, is at most 2, we find that the differences in the SSL and SL risk bounds is at most a constant factor of 2.

4.2.2 Bounds for the General Case

In the general case we do not assume that the target hypothesis is part of our hypothesis class. To still make use of the considered metric, the author proposes the following general recipe for bounds in that case. The starting point is the observation that bounds for randomized classifiers are generally tighter when compared to their deterministic counterparts [McAllester, 2003c, Langford and Shawe-Taylor, 2002]. The idea is now to use such a randomized classifier f_{rand} as an anchor, similarly to the target hypothesis in the realizable case. To get a bound for a classifier f we then can use the bound for the randomized classifier together with a slack term that includes $\hat{d}(f_{\text{rand}}, f)$. Depending on which kind of randomized classifier we take, we obtain different bounds. This includes for example PAC-Bayesian bounds as well as bounds based on cross-validation and bagging methods. They explicitly derive a cross-validation bound, where the randomized classifier is given by a uniform distribution over the classifiers obtained in the multiple cross-validation rounds.

4.3 Using Unlabeled Data to Combine Multiple Hypothesis Spaces

Leskes [2005] presents another scheme that relies on measuring the classification agreement between hypotheses on unlabeled data. The idea here is to use a boosting scheme, so we start with $L \in \mathbb{N}$ different hypothesis classes $H^1, ..., H^L$. We want to find the best fitting hypothesis over all L hypothesis classes $H^1, ..., H^L$. As that would generally lead to an overly increased complexity, the paper reduces the set of possible hypotheses by only considering those that agree sufficiently on the unlabeled data. In this context sufficiently means that we switch to a new hypothesis class H_v for a $v > 0$ that is defined as

$$H_v = \{(h^1, ..., h^L) \in H^1 \times ... \times H^L \mid V(h^1, ..., h^L) \leq v\},$$

where

$$V(h^1, ..., h^L) := \mathbb{E}_X[\frac{1}{L}\sum_i h^i(X)^2 - (\frac{1}{L}\sum_i h^i(X))^2].$$

The term $V(h^1, ..., h^L)$ essentially measures the variance of disagreement within L different hypotheses and is approximated with the unlabeled data. The hypothesis class H_v only keeps those collections of hypotheses that have a sufficiently small variance of disagreement. The paper then presents a generalization bound that holds for all h^l with $1 \leq l \leq L$ simultaneously and the bound depends on the maximum Rademacher complexity of the L base hypothesis classes $H^1, ..., H^L$.

5 Learning Under Weak Assumptions

In the previous two sections we investigated what is possible for semi-supervised learners when we do not have any additional assumptions. Now we investigate what a SSL can achieve under what we call *weak* assumptions. With weak assumptions we mean those that cannot essentially change the learning of $O(\frac{1}{\sqrt{n}})$, but rather gives improvements by a constant which can depend on the hypothesis class. In Section 6 we will investigate what we have to assume to escape the $\frac{1}{\sqrt{n}}$ regime. We first cover the work of Balcan and Blum [2010], as it is a general framework that allows us to analyze the learning guarantees for multiple semi-supervised learners. They show that semi-supervised learners that fall in this framework learn by a constant faster than supervised learners, where the constant depends on the hypothesis class and the semi-supervised learner we use.

We then cover in more detail the idea of co-training. Although co-training can also be viewed in the framework of Balcan and Blum [2010] we want to present a few more details on it. In particular we present the work of Sridharan and Kakade [2008] who formulate the assumption of co-training in an information theoretical framework, which allows to precisely quantify the bias-variance trade-off.

5.1 A General Framework to Encode Weak Assumptions

We start with the work done by Balcan and Blum [2010], as it offers an elegant way to formalize different assumptions in a general framework. Many existing methods can be cast in this framework; transductive support vector machines [Joachims, 1999, Boyd and Vandenberghe, 2004], Multi-View assumptions [Blum and Mitchell, 1998, Leskes, 2005, Sridharan and Kakade, 2008] and transductive graph-based methods [Blum and Chawla, 2001]. The idea is to introduce a function χ that measures the compatibility between a hypothesis h and the marginal distribution $P(X)$. Compatibility can mean many different things in this context. As a simple example we could call a hypothesis h compatible with a marginal distribution $P(X)$ if its decision boundary goes through low density regions. As we usually only observe a finite sample size, the function χ needs to be defined for each point in the feature space, so one sets

$$\chi : H \times \mathcal{X} \to [0, 1]. \tag{19}$$

The compatibility measure χ gives then rise to the function

$$R_{\text{unl}}(h) := 1 - \mathbb{E}_{X \sim P(X)}[\chi(h, X)], \tag{20}$$

which we will call the *unsupervised loss*. We will try to optimize it in addition to the loss measured on the labeled sample. The paper states several more theorems in the same flavor as the one presented here. The differences are mostly in the realizability assumptions (regarding the unsupervised and the supervised error) and the bounding technique. They present bounds derived from uniform convergence as well as bounds based on covering numbers. The following theorem is the double agnostic case (neither the labeled nor the unlabeled loss have to be zero).

Theorem 6 (Theorem 10)**.** *Let* $h_t^* = \arg\min_{h \in H}[R(h) \mid R_{\text{unl}}(h) \leq t]$*. Then, given an unlabeled sample size of at least*

$$\mathcal{O}\left(\frac{\max[VC(H), VC(\chi(H))}{\epsilon_2} \ln \frac{1}{\epsilon_2} + \frac{1}{\epsilon_2^2} \ln \frac{1}{\delta}\right)$$

we have that

$$m(h^{\text{SSL}}, H, \epsilon, \delta) \leq \frac{32}{\epsilon^2}\left[VC(H(t + 2\epsilon_2)) + \ln \frac{2}{\delta}\right], \tag{21}$$

where h^{SSL} is the hypothesis that minimizes $\hat{R}(h^{\mathrm{SSL}})$ subject to $\hat{R}_{\mathrm{unl}}(h^{\mathrm{SSL}}) \leq t+\epsilon$ and $H(t) := \{h \in H \mid R_{\mathrm{unl}}(h) \leq t\}$. Here $\hat{R}$ is the empirical risk measured with the sample S_n and $\hat{R}_{\mathrm{unl}}$ is the empirical unlabeled risk measured on the sample U_m.

We note that the original paper uses a different measure of complexity, so the term $VC(H(t + 2\epsilon_2))$ is different. We use the standard VC-dimension instead to avoid additional notation and to allow for an easier comparison to other results. They use a complexity notion that in Vapnik [1998] could be found under (the exponentiated) annealed entropy and has the advantage to be distribution dependent.

We now compare Theorem 6 to the results of the previous section, in particular to Conjecture 1 and the answers to this as found in Theorems 3 and 4. We know that in the purely supervised case we can achieve a similar sample complexity as (21) by replacing $VC(H(t + 2\epsilon_2))$ with $VC(H)$. As we know that the sample complexity given by (21) is tight up to constants (compare Chapter 6 from Shalev-Shwartz and Ben-David [2014]), we know that the sample complexity between a purely supervised learner and the semi-supervised learner as defined in this paper cannot differ by more than $\mathcal{O}\left(\frac{VC(H)}{VC(H(t+2\epsilon_2))}\right)$. So the gap in the learning rates is indeed given by a constant that only depends on the hypothesis class as postulated by Conjecture 2. This constant can, however, be infinite if $VC(H)$ is infinite but $VC(H(t + 2\epsilon_2))$ is finite. This is exactly the type of example that refuted the conjecture and which we presented in Section 3.2.

Theorem 6 quantifies to some degree the fundamental bias-variance trade-off in SSL when we use assumptions. Employing a semi-supervised compatibility function we reduce the variance of the training procedure as we effectively restrict the original hypothesis space H. If, however, the compatibility function does not match the underlying problem, we bias the procedure away from good solutions.

5.2 Assuming that the Feature Space can be Split

In multi-view learning, also sometimes called co-regularization or co-training, one assumes that the feature space $\mathcal{X}$ can be decomposed as $\mathcal{X} = \mathcal{X}^1 \times \mathcal{X}^2$, and each partial feature space $\mathcal{X}^1, \mathcal{X}^2$ is already enough to learn. In the early work on co-training Blum and Mitchell [1998] use the idea in a web page classification set. One part of the features, say $\mathcal{X}^1$, is given by the text on the web page itself, while the other one, $\mathcal{X}^2$, is given by the anchor text of hyperlinks pointing to the web page. The idea is that if both partial features spaces have sufficient information about the correct label, we would expect that a correct classifier predicts the same label given any of the two partial features. We can thus discard classifiers that disagree on the two views.

There are multiple theoretical results about this approach, it can be for example analyzed in the framework of the previous section. Rosenberg and Bartlett [2007] and Farquhar et al. [2006] analyze a Rademacher complexity term under the multi-view assumption. Sindhwani and Rosenberg [2008] define a kernel that directly includes the assumption as a regularization term, and thus find a RKHS where co-regularization automatically happens.

Here we detail the work of Sridharan and Kakade [2008], as this ties in best with the other results we present. In addition their information theoretic framework allows to also analyze the penalty one suffers if the assumption is not exactly true. We split the random variable X which takes values in $\mathcal{X}$ into $X = (X^1, X^2)$. In their framework the multi-view assumption can be formalized as follows.

Multi-View Assumption Let $I(A; B \mid C)$ be the mutual information between random variables A and B, conditioned on knowing already the random variable C. Then there exists an ϵ_{info} such that

$$I(Y; X^2 \mid X^1) \leq \epsilon_{\text{info}} \tag{22}$$

and

$$I(Y; X^1 \mid X^2) \leq \epsilon_{\text{info}}. \tag{23}$$

Intuitively this states that once we know one of the features, the other feature will not tell us much more about Y.

Comparing this to co-training we can see it as a relaxation. In co-training one assumes that each view is already sufficient to fully learn, which corresponds here to $\epsilon_{\text{info}} = 0$. If, however, $\epsilon_{\text{info}} > 0$, we cannot learn perfectly from one view. (But this is fine in this framework). We assume then, that we have for each view X^1 and X^2 a corresponding hypothesis set H^1 and H^2. We will do predictions with *pairs* of hypotheses $(f_1, f_2) \in H^1 \times H^2$. The paper uses the notion of compatibility functions (19). In particular they define a compatibility function $\chi : H := H^1 \times H^2 \to [0, 1]$ as $\chi(h^1, h^2, x) := d(f_1(x^1), f_2(x^2))$, where $d : \mathcal{Y} \times \mathcal{Y} \to [0, 1]$ is some sort of distance measure that fulfills a relaxed triangle inequality and $x = (x^1, x^2)$ is a sample. The distance d measures in essence how much f_1 and f_2 agree on a sample x. For a given threshold $t \in \mathbb{R}$ we find now the best *pair* of hypotheses with the constrained empirical risk minimization problem

$$\min_{(h^1,h^2)\in H} \quad \sum_{i=1}^{n} l(h^1(x_i^1), y_i) + l(h^2(x_i^2), y_i) \quad \text{subject to} \quad \hat{R}_{\text{unl}}(h^1, h^2) \le t. \tag{24}$$

Recall the definition of $R_{\text{unl}}(h)$ from Equation (20). The main theorem, which gives guarantees on the solution found by the procedure above, needs the following notation. Let β_*, β_*^1 and β_*^2 be the Bayes error, measured with the loss l, when learning from $X^1 \times X^2$, X^1 and X^2 respectively. We also set $\epsilon_{\text{regret}} = \max\{R(f_*^1) - \beta_*^1, R(f_*^2) - \beta_*^2\}$, where f_*^i is the best predictor from H^i. Finally we set $\hat{H}(t) = \{(h^1, h^2) \in H \mid \hat{R}_{\text{unl}}(h^1, h^2) \le t\}$.

Theorem 7. *Assume that the loss l is bounded by 1. There exists an $t \in \mathbb{R}$ (depending among other on ϵ_{info}, ϵ_{bayes} and m), such that under some further regularity conditions on $\chi = d$ and the loss l, and given at least $m(\hat{H}(t), \epsilon, \delta)$ labeled samples, with probability $1 - \delta$*

$$\frac{R(\hat{h}^1) + R(\hat{h}^2)}{2} \le \beta_* + \epsilon + \epsilon_{\text{bayes}} + \sqrt{\epsilon_{\text{info}}}. \tag{25}$$

We see now that the information theoretic assumption allows us to explicitly describe the bias introduced when switching from the full hypothesis set H to the restricted one $\hat{H}(t)$. This bias is given by $\sqrt{\epsilon_{\text{info}}}$.

6 Learning Under Strong Assumptions

In the previous section we analyzed assumptions that only could give us a constant improvement, and did not allow us to escape the general learning rate of $\frac{1}{\sqrt{n}}$. Now we analyze assumptions which allow us to escape this regime, and can even give exponentially fast convergence. The following example illustrates the basic idea behind that. Assume we are given a set of unlabeled data and we use it to cluster the data. If we assume that the clustering is correct, meaning that each cluster corresponds to a class, we essentially need only enough labeled data to identify which cluster belongs to which class. The work we present in this section extends this idea in various ways and answers the following questions. What if we have class overlap? What if there is noise in the clusters? What about regression?

6.1 Assuming that the Model is Identifiable

One of the classic works in semi-supervised learning, that deals with a topic closely related to sample complexity, was done by Castelli and Cover [1995]. The setting is very restricted but can give exponentially fast convergence rates to the Bayes risk in the number of labeled samples n. This is very powerful considering that the results of the previous sections could often not essentially fasten the rate of $\frac{1}{\sqrt{n}}$ (compare for example Inequality (21) after solving for ϵ).

The first key assumption to obtain those results lies in the data generation process. First the label is drawn with $P(y = 1) = \eta$ and $P(y = 0) = \bar{\eta}$ and then a feature is drawn according to a density $f_y(x)$. Unlabeled data is thus drawn from the mixture $\eta f_1 + \bar{\eta} f_2$. The second key assumption is that the class of mixture models is identifiable, i.e. that we can infer the mixture model uniquely given only unlabeled data. After observing enough unlabeled data to identify the mixture we only have to figure out how to label each part of the two mixture components. As we thus have only to decide between two alternatives we can find a classifier h by a simple likelihood ratio test, which converges exponentially fast to the Bayes risk in the number of the labeled samples n:

$$R(h) - \min_{h\in H} R(h) \le \exp\left(n \ln(2\sqrt{\mu\bar{\mu}} \int \sqrt{f_1(x) f_2(x) dx}) + o(n)\right) \tag{26}$$

For the analysis it is necessary to assume that one has an infinite amount of unlabeled data. The work is continued in Castelli and Cover [1996], where the authors consider cases where we already have knowledge about the densities f_y. Sinha and Belkin [2007] extend a similar framework to the case where the marginal distribution $P(x)$ is unknown. They assume instead that $P(x)$ can be well estimated with a mixture of two spherical Gaussian distributions with density functions $f_1(x)$ and $f_{-1}(x)$. In particular they assume that $||f_1 - P(\cdot|Y = 1)||_S$ and $||f_{-1} - P(\cdot|y = -1)||_S$ can be bounded with a small number, where $|| \cdot ||_S$ is a Sobolev norm. Finally we want to mention the work of Ratsaby and Venkatesh [1995], where exponential decay of excess risk is achieved under the assumptions of well-specification and the model class are mixtures of two spherical Gaussian distributions.

6.2 Assuming that Classes are Clustered and Separated

In Rigollet [2007] we are presented explicit bounds on the generalization error using another formulation of the cluster assumption. It closely resembles the work of the previous section and under their assumption we again obtain exponentially fast convergence. Their first and simple setup is that we are given a collection of pairwise disjoint clusters $C_1, C_2, ...$ and we make a *cluster assumption*, i.e we assume that the labeling function $x \mapsto \text{sign}(P(Y = 1 \mid X = x) - \frac{1}{2})$ is constant on each cluster C_i. So the clusters have a label-purity of some degree, which we can specify by

$$\delta_i = \int_{C_i} |2P(Y = 1|X = x) - 1| dP(x), \tag{27}$$

where the cluster C_i is pure iff δ_i is either 1 or 0. Assuming that we know the clusters, we let $h_n^{\text{SSL}}(x)$ be the majority voting classifier per cluster. More formally, given a labeled sample S_n let $X_i^+ := \{(x, y) \in S_n \mid x \in C_i, y = 1\}$ and similarly $X_i^- := \{(x, y) \in S_n \mid x \in C_i, y = -1\}$. Then given a new data point $x \in C_i$ we set

$$h^{\text{SSL}}(x) = \begin{cases} 1 & \text{if } |X_i^+| \geq |X_i^-| \\ -1 & \text{if } |X_i^+| < |X_i^-|. \end{cases} \tag{28}$$

Note that this defines only a function on the clusters. The paper argues, however, that unlabeled data cannot help where no unlabeled data was observed. Consequently it only analyses the possible gain from unlabeled data on the clusters. Thus the excess risk is now restricted to the set $C := \bigcup C_i$, so we set the excess risk as

$$\mathcal{E}_C(h) = \int_C |2P(Y = 1|X = x) - 1| I_{\{h(x) \neq h^*(x)\}} dP(x),$$

where h^* is the Bayes classifier. The following theorem describes the gain one can make with respect to the expected cluster excess risk.

Theorem 8 (Theorem 3.1)**.** *Let $(C_i)_{i \in I}$ be a collection of sets with $C_i \subset \mathcal{X}$ for all $i \in I$ such that this collection fulfills the above defined cluster assumption. Then the majority voting classifier h_n^{SSL} as defined above satisfies*

$$\mathbb{E}_{S_n, U_m} \left[\mathcal{E}_C(h_n^{\text{SSL}}) \right] \leq 2 \sum_{i \in I} \delta_i e^{\frac{-n\delta_i^2}{2}}. \tag{29}$$

So knowing the clusters we recover the exponential convergence in the labeled sample size as in Castelli and Cover [1995]. The biggest effort of the paper goes, however, in the definition of clusters and the finite sample size estimation of such. The derivations are rather long and here we limit ourselves to describe the underlying intuition. First we assume that the marginal distribution $P(X)$ allows for a density function $p(x)$ with respect to the Lebesgue measure. With that one can define the density level sets of $\mathcal{X}$ w.r.t. a parameter $\lambda > 0$ as $\Gamma(\lambda) := \{x \in \mathcal{X} \mid p(x) \geq \lambda\}$. For a fixed $\lambda > 0$ we think of a clustering essentially as path-connected components of the density level sets $\Gamma(\lambda)$, where it is ensured that pathological cases are excluded. Estimating the set $\Gamma(\lambda)$ with finitely many unlabeled samples adds a slack term to Inequality (29) that drops polynomially in the unlabeled sample size. So, to ensure that we still can learn exponentially fast, the number of unlabeled samples has to grow exponentially with the number of labeled samples.

6.3 Assuming that the Classes are Clustered but Not Necessarily Separated

Singh et al. [2008] propose a different formalization of the cluster assumption, one that allows to distinguish cases where SSL does help and where not. This is done by restricting the class of distributions $\mathcal{P}$ and then investigating which of those distributions allow for successful semi-supervised learning. The class $\mathcal{P}$ is constructed such that the marginal distributions are constituted of different clusters that are sometimes easy to distinguish and sometimes not. The marginal densities $p(x)$ from $\mathcal{P}$ are given by mixtures of K densities p_k. So $p(x) = \sum_{i=1}^K a_k p_k(x)$ with $a_k > 0$ and $\sum_{i=1}^K a_k = 1$ and each p_k has support on a set $C_k \subset \mathcal{X}$ which fulfills some regularity conditions. We call the sets C_k clusters, and each one is assumed to have its own smooth label distribution function $p_k(y \mid x)$. So with probability a_k we draw from $p_k(x)$ and then label x according to $p_k(y \mid x)$. We further only consider distributions that lead to clusters with margin, with our without overlap, of at least γ (see also Figure 5), and denote the resulting class of distributions by $\mathcal{P}(\gamma)$. In this formulation the clusters are not of the main interest, but rather what the authors call the *decision sets*.

To define a decision set we denote with C_k^c the complement of C_k and define $C_k^{\neg c} := C_k$. A set $D \subset \mathcal{X}$ is called a decision set if it can be written as

$$D = \bigcap_{k \in K} C_k^{i_k}$$

for $i_k \in \{c, \neg c\}$, see Figure 5 (b) for an example. The advantage of the decision sets over the clusters are that the full distribution $p(x, y)$ is not necessarily smooth on each cluster, as they might exhibit jumps at the borders. On the decision sets, however, $p(x, y)$ will be smooth, if each $p_k(y \mid x)$ is smooth. Thus, if we would know the decision sets we could use a semi-supervised learner that uses the smoothness assumption.

The main theorem answers the question whether or not one can learn the decision sets from finitely many unlabeled points.

Theorem 9 (Corollary 1)**.** *Let $\mathcal{E}(h) = R(h) - R^*$ be the excess risk with respect to the Bayes classifier R^*. Assume that $\mathcal{E}$ is bounded and that there is a learner h_n^D that has knowledge of all decision sets D and fulfills the following excess risk bound.*

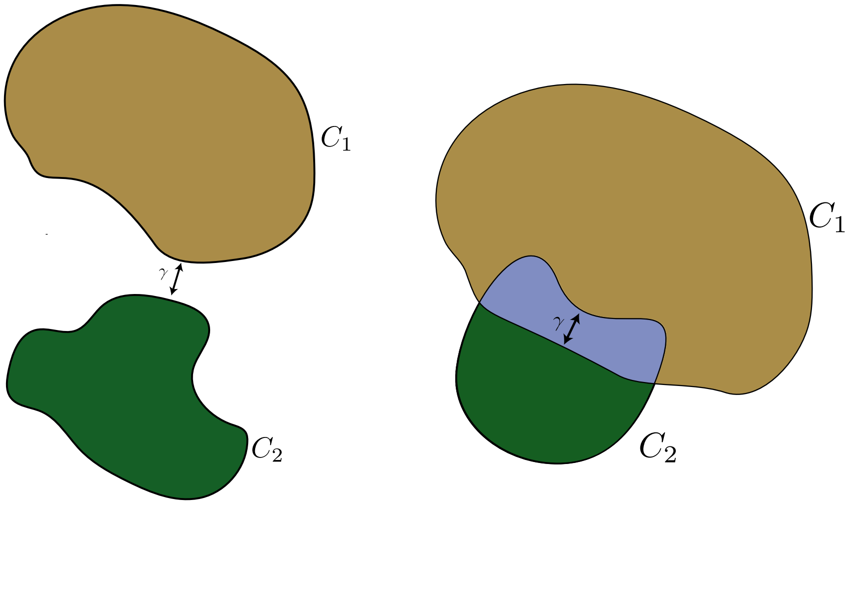
$$\sup_{P \in \mathcal{P}(\gamma)} \mathbb{E}_P[\mathcal{E}(h_n^D)] \leq \epsilon_2(n) \tag{30}$$

Assume that $|\gamma| > 6\sqrt{d}\kappa_0(\frac{(\ln m)^2}{m})^{\frac{1}{d}}$, where κ_0 is a constant, then a semi-supervised learner $h_{n,m}^{\mathrm{SSL}}$ exists such that

$$\sup_{P \in \mathcal{P}(\gamma)} \mathbb{E}_P[\mathcal{E}(h_{n,m}^{SS})] \leq \epsilon_2(n) + O\left(\frac{1}{m} + n\left(\frac{(\ln m)^2}{m}\right)^{\frac{1}{d}}\right). \tag{31}$$

Note the following. If the learner h_n^D that knows the decision sets has a convergence rate of $\epsilon_2(n)$, it follows from Inequality (31) that the unlabeled data needs to increase with a rate of $\epsilon_2(\frac{1}{n})$ to ensure that the semi-supervised learner has the same convergence rate as h_n^D. For example, if h_n^D converges exponentially fast, we need a exponentially much more unlabeled then labeled data, which is the same finding as in the previous section.

The intuition here is fairly simple. The bigger γ the less unlabeled samples we need to estimate the decision sets D, and once we know those, we can perform as well as h_n^D. To analyze if a semi-supervised learner that first learns the decision sets empirically has an advantage over all supervised learners, they first find minimax lower bounds for all fully supervised learners. They then give upper bounds for a specific semi-supervised learner and the conclusions are intuitive: For SSL to be useful, the parameter γ and the number of unlabeled samples should be such that the fully supervised learner cannot distinguish the decision sets, while the semi-supervised learner can. So γ should not be too big, as then the supervised learner can also distinguish the decision sets. And, of course, the unlabeled data should not be too little, as then the semi-supervised learner cannot distinguish the decision sets.



(a) The clusters C_1 and C_2 are separated with margin γ. The different decision regions are here just the clusters.

(b) The clusters C_1 and C_2 are have an overlap (light blue) with margin γ. The three colors also constitute three different decision sets.

Figure 5: Picture (a) shows the concept of a positive γ-margin, while (b) shows a negative γ-margin.

To present specific differences between SSL and SL the authors assume that $\mathcal{X} = [0,1]^d$ and that the conditional expectations $\mathbb{E}_{Y\sim p_k(Y|X=x)}[Y|X=x]$ are Hölder-α smooth functions in x. Depending on γ the paper presents a table for cases when SSL can be essentially faster than SL. In those cases the SL has an expected lower bound for the convergence rate of $n^{-\frac{1}{d}}$ while the convergence rate of the SSL is upper bounded by $n^{-\frac{2\alpha}{2\alpha+d}}$.

6.4 Assuming the Regression Function is Smooth Along A manifold

As we will elaborate further in the discussion section, an issue in SSL is that most methods are based on assumptions on the full distribution. The problem is that we usually cannot verify whether the assumptions holds or not. This is crucial to know, since in case the assumption does not hold, it is quite likely that we want to use a supervised learner instead. The work of Azizyan et al. [2012] is one of the few papers that touches on that topic as they introduce a semi-supervised learner that depends on a parameter α, where $\alpha = 0$ recovers a purely supervised learner. The paper then gives generalization bounds for the semi-supervised learner when we cross-validate α. As this work uses the regression setting, while most other presented papers deal with classification, and gives a clean formalization of the SSL, we present here the details. The authors use a version of the manifold assumption, so we enforce our estimated regression function $h^{\text{SSL}}(x)$ to behave smoothly in high density regions. The density of the marginal distribution $P(X)$ is measured with a smoothed density function $p_\sigma(x)$

$$p_\sigma(x) := \int \frac{1}{\sigma^d} K\left(\frac{||x-u||}{\sigma}\right) dP(u), \tag{32}$$

where K is a symmetric kernel on $\mathbb{R}^d$ with compact support and $\sigma > 0$. Let $\Gamma(x_1, x_2)$ be the set of all continuous paths $\gamma : [0, L(\gamma)] \to \mathbb{R}^d$ from $x_1 \in \mathbb{R}$ to $x_2 \in \mathbb{R}$ with unit speed and where $L(\gamma)$ is the length of γ. With this we can define a new metric (the so-called exponential metric) on $\mathbb{R}^d$ that depends on a parameter $\alpha \geq 0$ and the smoothed density $p_\sigma(x)$.

$$D(x_1, x_2) = \inf_{\gamma\in\Gamma} \int_0^{L(\gamma)} e^{-\alpha p_\sigma(\gamma(t))} dt \tag{33}$$

First note that $\alpha = 0$ corresponds to the Euclidean distance. Second, note that high values of $p_\sigma(x)$ on the path between two points x_1 and x_2 lead to shorter distances between those points in the new metric, and this is emphasized with large α. If we assume that Q is another kernel and we set $Q_\tau(x) := \frac{1}{\tau^d} Q(\frac{x}{\tau})$ we can define the semi-supervised estimator as

$$h^{\text{SSL}}(x) := \frac{\sum_{i=1}^n y_i Q_\tau(\hat{D}(x, x_i))}{\sum_{i=1}^n Q_\tau(\hat{D}(x, x_i))}. \tag{34}$$

The estimator is thus a nearest-neighbor regressor, where neighbors are weighted according to their distance in the D-metric. The following theorems gives bounds on the squared risk of h^{SSL} under the assumption that $\sup\limits_{y\in\mathcal{Y}} |y| = M < \infty$.

Theorem 10 (Theorem 4.1). *Let $\mathcal{P}(\alpha, \sigma, L)$ be a class of probability measures that fulfill certain regularities depending on parameters $\alpha, \sigma, L \geq 0$ (more details after the Theorem). Assume that for all $P \in \mathcal{P}$ we have $P(||\hat{p}_\sigma - p_\sigma|| \geq \epsilon_m) \leq \frac{1}{m}$, then*

$$\mathbb{E}_{S_n, U_m}[R(h^{\text{SSL}}] \leq L^2(\tau e^{\alpha\epsilon_m})^2 + \frac{1}{n} M^2 (2 + \frac{1}{e}) \mathcal{N}_{P,\alpha,\sigma}(e^{-\alpha\epsilon_m}\frac{\tau}{2}) + \frac{4M^2}{m}). \tag{35}$$

In this notation $\mathcal{N}_{P,\alpha,\sigma}(\epsilon)$ is the *covering number* of P in the D-metric: The minimum number of closed balls in $\mathcal{X}$ of size ϵ w.r.t to the D-metric necessary to cover the support of $P(X)$, see also Shalev-Shwartz and Ben-David [2014, Chapter 27]. In the Euclidean case, so when $\alpha = 0$, we can bound $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \leq (\frac{C}{\epsilon})^d$ with a constant C. The covering number can be much smaller when $\alpha > 0$ and $P(X)$ is concentrated on a manifold with dimension smaller than d. The regularity conditions on $\mathcal{P}(\alpha, \sigma, L)$ are essentially the following. First we assume that $P(X)$ has compact support. Second, all regression functions $f_P(x) = \mathbb{E}P(Y \mid X = x) : \mathbb{R}^d \to \mathbb{R}$ are L-Lipschitz continuous, where the domain $\mathbb{R}^d$ is equipped with the exponential metric D and the co-domain $\mathbb{R}$ is equipped with the Euclidean distance.

As the previous Theorem might be quite difficult to parse, the paper offers a simplified corollary, under some further regularity conditions.

Corollary 1 (Corollary 4.2). *Assume that $\mathcal{N}_{P,\alpha,\sigma}(\delta) \leq (\frac{C}{\delta})^\xi$ for some certain range of δ. Furthermore assume that m is large enough and that $\tau(n, \alpha, \epsilon_m, \xi)$ is well chosen. Then for all $P \in \mathcal{P}(\alpha, \sigma, L)$*

$$\mathbb{E}_{S_n, U_m}[R(h^{\text{SSL}}] \leq \left(\frac{C}{n}\right)^{\frac{2}{2+\xi}}. \tag{36}$$

The paper then analyzes the additional penalty we occur in trying to find the best α. This is done by discretizing the parameter space $\Theta = \mathcal{T} \times \mathcal{A} \times \Sigma$ such that $\theta = (\tau, \alpha, \sigma) \in \Theta$ and $|\Theta| = J < \infty$. Assume that we have in addition to the training sample S_n also a validation set $V = \{(v_1, z_1), ..., (v_n, z_n)\}$, for convenience also of size n. Let h_θ^{SSL} be the semi-supervised hypothesis trained on S_n with the parameters θ. We then choose the final hypothesis h^{SSL} by choosing θ with cross-validation

$$h^{\text{SSL}} := \arg\min_{h_\theta^{\text{SSL}}} \sum_{i=1}^{n} (h_\theta^{\text{SSL}}(v_i) - z_i)^2. \tag{37}$$

Theorem 11 (Theorem 6.1)**.** *Let $\mathcal{E}(h) := R(h) - R(h^*)$ be the excess risk, where h^* is the true regression function. There are constants [not universal, depend to some degree on the problem] $0 < a < 1$ and $0 < t < \frac{15}{38(M^2+\sigma^2)}$ such that*

$$\mathbb{E}_{S_n,U_m,V}[\mathcal{E}(h^{\text{SSL}})] \leq \frac{1}{1-a}\left(\min_{\theta\in\Theta} \mathbb{E}_{S_n,U_m}[\mathcal{E}(h_\theta^{\text{SSL}})] + \frac{\ln(nt4M^2) + t(1-a)}{nt}\right), \tag{38}$$

This is particularly interesting since we implicitly compare to the supervised solution, as long as we include $\alpha = 0 \in \mathcal{A}$. From Inequality (38) we see that the validation process introduces a penalty term of $O(\frac{\ln(n)}{n})$. In the worst case this can be seen as an additional error term if we use the semi-supervised method, but the assumption is actually not true.

Finally the authors identify a case where the semi-supervised learning rate can be strictly better than the supervised learning rate, much like we have seen in Section 3.2. In particular, they construct a set of distributions $\mathcal{P}_n$, which depends on the number of labeled samples, such that

1. the estimator $h^{\text{SSL}}(x)_{\tau,\alpha,\sigma}$, as defined in Equation (34), fulfills

   $$\sup_{P\in\mathcal{P}_n} \mathbb{E}_{S_n}[R(\hat{f}_{\tau,\alpha,\sigma})] \leq \left(\frac{C}{n}\right)^{\frac{2}{2+\xi}},$$

   under the assumption that $m \geq 2^{\frac{2}{2+\xi}}$.
2. for all purely supervised estimators h^{SL} we have that

   $$\sup_{P\in\mathcal{P}_n} \mathbb{E}_{S_n}[h^{\text{SL}}] \geq \left(\frac{C}{n}\right)^{\frac{2}{d-1}}.$$

To obtain essentially different learning rates we need that $\xi < d - 3$, which is the case if P is concentrated on a set with dimension strictly less than $d - 3$ [Azizyan et al., 2012, Lemma 1]. It is also worth noting that the construction of $\mathcal{P}_n$ works by concentrating the distributions more for bigger n. If $\mathcal{P}_n$ does not concentrate, and remains smooth for bigger n, the labeled data is already enough to approximate the marginal distribution.

This is similar to the work presented in Section 6.3, as they also show that SSL can only work if the marginal distribution $P(X)$ is not too easy to identify. We can also draw parallels to the work presented in Section 3.2.3; if we would restrict the domain distributions such that only smooth circle embeddings would be allowed, a supervised learner could also learn efficiently. This is because then a finite number of labeled samples would be sufficient to learn the domain distribution uniformly, so the semi-supervised learner would loose its benefits.

7 Learning in the Transductive Case

While many methods use unlabeled data to find better classification rules, some consider schemes where one only cares about the labels of the unlabeled data. Those methods are often called transductive [Vapnik, 1998, Chapter 8]. We present the most important theoretical results. A more detailed survey on theoretical and practical transductive learning can be found in Chapter 2 of Pechyony [2008]. In Subsection 7.1 we present learning bounds in the transductive case. They often arise as direct extension of the inductive case and related concepts. In Subsection 7.2, which cannot be found as part of Pechyony [2008], we present two papers that touch on the topic of so-called safe semi-supervised learners. Their aim is to construct semi-supervised learners that are never worse than their supervised counterparts.

One can distinguish two transductive settings, where the essential difference is that in one setting we sample without replacement, so the samples become dependent. The work about transductive learning which we present here deals with Setting 1, mostly because of convenience. We note, however, that one can transform bounds from Setting 1 to bounds from Setting 2 [Vapnik, 1998, Theorem 8.1].

Setting 1

1. We start with a fixed set of points $X_{n+m} = \{x_1, ..., x_{n+m}\}$.
2. We reveal the labels Y_n of a set $X_n \subset X_{n+m}$ which is uniformly selected at random. For notational convenience we usually assume w.l.o.g that X_n are the first n and X_m are the last m points of X_{n+m}.
3. Based on $S_n = (X_n, Y_n)$ and X_m we try to find a classifier h with good performance on $R_m(h) := \sum_{i=n}^{n+m} l(x_i, y_i)$.

Setting 2

1. We start with a fixed distribution P on $\mathcal{X} \times \mathcal{Y}$.
2. We draw n i.i.d. samples according to P to obtain a training set S_n. We draw m i.i.d. samples according to $P(X)$ to obtain a test set X_m.
3. Based on $S_n = (X_n, Y_n)$ and X_m we try to find a classifier h with good performance on $\mathbb{E}_{S_n, X_m}\left[\frac{1}{m}\sum_{i=n}^{n+m} l(h(x_i), y_i)\right]$.

Note that in this section our test error is denoted by $R_m(h)$ and the training error by $R_n(h)$. This reflects that the test is of size m while the training set of size n. We will not use the hat notation here, as in the transductive setting we do not necessarily have an underlying distribution.

7.1 Transductive Learning Bounds

7.1.1 Vapnik's Implicit Transductive Bound

Transductive inference goes back to Vapnik [1982]. We present the result found as Equation (8.15) in Theorem 8.2. from Vapnik [1998]. Assume that we are given $n + m$ samples and we pick at random n samples on which we can train. We then want to estimate the error we make on the leftover m samples. Vapnik shows that a hypergeometric distribution describes the probability that the observed error on the train and test set is bigger than ϵ

$$P\left(\frac{|R_m(h) - R_n(h)|}{\sqrt{R_{n+m}(h)}} > \epsilon\right).$$

Let ϵ^* be the smallest $\epsilon > 0$ such that

$$P\left(\frac{|R_m(h) - R_n(h)|}{\sqrt{R_{n+m}(h)}} > \epsilon\}\right) \leq 1 - \delta.$$

Using a uniform bound[9] and substituting $R_{n+m} = \frac{m}{n+m}R_m + \frac{n}{n+m}R_n$ one can derive the following result.

Theorem 12. *For all $h \in \{-1, 1\}^{n+m}$ the following inequality holds with a probability of $1 - \delta$*

$$R_m(h) \leq R(h) + \frac{(\epsilon^*)^2 m}{2(m+n)} + \epsilon^* \sqrt{R(h) + \left(\frac{\epsilon^* m}{2(m+n)}\right)^2} \qquad (39)$$

The problem of this inequality is that the term ϵ^* is an implicit function of n, m, δ and h and thus it is unclear what the learning rates are that we can actually achieve. This problem is addressed in the paper presented in the next section.

7.1.2 Bounds as a Direct Extension of Inductive Bounds

The transductive bound of Inequality (39) is difficult to interpret as it contains a function which can only be implicitly calculated. Derbeko et al. [2011] find explicit transductive bounds in a PAC-Bayes framework. We present a bound from the paper which is essentially a direct extension of an inductive bound from McAllester [2003a]. To present the result they use a Gibbs classifier. For that, let q be any distribution over the hypothesis set H. The Gibbs classifier G_q classifies a new instance $x \in \mathcal{X}$ with an $h \in H$ drawn accordingly to q. The risk of G_q over the set S_n is then $R_n(G_q) = \mathbb{E}_{h\sim q}[\frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)]$.

[9] Note that in the transductive case we effectively can have only finitely many different hypotheses.

Theorem 13 (Theorem 17)**.** *Let p be any (prior) distribution on H, which may depend on S_{n+m}, and let $\delta > 0$. Then for any randomly selected subset $S_n \subset S_{n+m}$ and for any distribution q on H, it holds with probability at least $1 - \delta$ that*

$$R_m(G_p) \leq R_n(G_p) + \frac{m+n}{m}\left(\sqrt{\frac{2R_n(G_p)(\mathrm{KL}(q||p) + \ln\frac{n}{\delta})}{n-1}} + \frac{2(\mathrm{KL}(q||p) + \ln\frac{n}{\delta})}{n-1}\right). \tag{40}$$

This theorem is indeed a direct extension of the inductive supervised case as found under Equation (6) in McAllester [2003a], the only difference is that the term $\frac{m+n}{m}$ is missing. Although McAllester [2003b] showed that under certain conditions one can select the prior p after having seen S_m, this is generally not allowed in inductive PAC-Bayesian theory. In the transductive setting this is allowed, as we only care about the performance on the points from the set S_{n+m}. In a way this is the same as learning with a fixed distribution when our fixed distribution has only mass on finitely many points [Benedek and Itai, 1991] .

Derbeko et al. [2011] exploit this by choosing a prior p with a cluster method. More precisely, after observing the dataset X_{n+m} one constructs c different clusterings on it. Each clustering leads to multiple classifiers by assigning all points in a cluster to the same class. One then puts essentially a uniform prior p on those classifiers and we select a posterior distribution q over the classifiers by minimizing Inequality (40), and obtain the Gibbs classifier G_q.

Comparing this approach to the fully supervised (and thus necessarily inductive) case, we realize that the possible performance improvements have the same flavor as the improvements one can gain in semi-supervised learning with assumptions, as analyzed in Sections 5 and 6. Using the clustering approach from above will reduce the penalty in Inequality (40) which is coming from $\mathrm{KL}(q||p)$. In other words: We reduce the variance of the classifier. On the other hand, using a clustering approach will bias our solution, and we will degrade over a supervised solution if clusterings have a high impurity, meaning that the clusterings don't have clear majority classes.

7.1.3 Bounds Based on Stability

In El-Yaniv and Pechyony [2006] transductive bounds are explored under the notion of stability, the assumption that the output of a classifier does not change much if we perturb the input a bit. The transductive bounds are an extension of the inductive bounds that use the notion of *uniform stability* [Bousquet and Elisseeff, 2002] and *weak stability* [Kutin and Niyogi, 2013, Kutin, 2002]. We present the simpler transductive bound based on uniform stability and explain the difference to weak stability.

Assume that $h^{\text{trans}} \in H$ is a transductive learner, so a hypothesis that we (deterministically) choose based on a labeled set S_n and an unlabeled set X_m. Furthermore define $S_n^{ij} := (S_n \setminus \{(x_i, y_i)\}) \cup \{(x_j, y_j)\}$ and $X_m^{ij} := (X_m \setminus \{x_j\}) \cup \{x_i\}$. So S_n^{ij} is the set we obtain when we replace in S_n the i-th example from the training set with the j-th example from the test set. We say that h^{trans} is *β-uniformly stable* if for all choices $S_n \subset S_{n+m}$ and for all $1 \leq i, j \leq n+m$ such that $(x_i, y_i) \in S_n$ and $x_j \in X_m$ it holds that

$$\max_{1 \leq k \leq n+m} |h^{\text{trans}}_{(S_n, X_m)}(x_k) - h^{\text{trans}}_{(S_n^{ij}, X_m^{ij})}(x_k)| \leq \beta. \tag{41}$$

In words: The transductive learner h^{trans} is β-uniformly stable if the output changes less than β if we exchange two points from the train and test set. The bounds are formulated using a γ-margin loss. For $\gamma > 0$ we set

$$l_\gamma(y_1, y_2) = \max(0, \min(1, 1 - \frac{y_1 y_2}{\gamma})). \tag{42}$$

Consequently we write $R_\gamma(h)$ for the risk of h when measured with the loss l_γ. Note that for $\gamma \to 0$ the l_γ loss converges to the $0-1$ loss.

Theorem 14 (Theorem 1)**.** *Let h^{trans} be a β-uniformly stable transductive learner and $\gamma, \delta > 0$. Then, with probability of at least $1 - \delta$ over all train and test partitions, we have that*

$$R_m(h^{\text{trans}}) \leq R_n^\gamma(h^{\text{trans}}) + \frac{1}{\gamma} O\left(\beta \sqrt{\frac{mn \ln\frac{1}{\delta}}{m+n}}\right) + O\left(\sqrt{(\frac{1}{m} + \frac{1}{n}) \ln\frac{1}{\delta}}\right). \tag{43}$$

Note that β will depend on n and m, and we would expect that the bigger our training set is, the less our algorithm changes if we exchange two samples from the train and test set. In the transductive bounds based on Rademacher complexities, in the section further below, one can achieve a convergence rate of $\frac{1}{\sqrt{\min(m,n)}}$. To obtain the same rate with Inequality (43) we need that β behaves as $O\left(\sqrt{(\frac{1}{n} + \frac{1}{m})\frac{1}{\min(n,m)}}\right)$. This stability rate can be indeed achieved for regularized RKHS methods as demonstrated by Johnson and Zhang [2007].

7.1.4 Bounds Based on Transductive Rademacher Complexities

Rademacher complexities are a well studied and established tool for risk bounds in the inductive case [Bartlett et al., 2005]. El-Yaniv and Pechyony [2009] introduce a transductive version of these quantities. While in the inductive case we have to chose our hypothesis class before seeing any data, the transductive case allows us to chose the hypothesis class data-dependent. The definition of the transductive Rademacher complexity of a hypothesis class H follows closely the inductive case and will be denoted by $\mathrm{tRad}(H)$. Utilizing the γ-margin loss function (42) and the corresponding empirical risk $R^{\gamma}(h)$, the paper shows then that for all $h \in H$

$$R_m(h) \leq R_n^{\gamma}(h) + \frac{\mathrm{tRad}(H)}{\gamma} + O\left(\frac{1}{\sqrt{\min(m,n)}}\right).$$

Examining the inequality on first sight, it seems somewhat surprising that the labeled and unlabeled data play an equivalent role in terms of convergence. While slow convergence for $n \ll m$ is not really surprising one has to realize that in the case where $m \ll n$ the transductive risk has a very high variance and thus we have large intervals for high-confidence estimations. This bound can be used to directly estimate the trandsuctive risk for transductive algorithms.

Maximov et al. [2016] make different use of Rademacher complexities to derive risk bounds for a specific multi-class algorithm. Their algorithm uses a given clustering based on the full data to find a hypothesis which is in some way compatible with the found clustering. The transductive multi-class Rademacher complexities then make direct use of this clustering. With this algorithm the authors show that if we have K initial classes one can achieve a learning rate in the order of $\tilde{O}(\frac{\sqrt{K}}{\sqrt{n}} + \frac{K^{3/2}}{\sqrt{m}})$ [Maximov et al., 2016, Corollary 4]. Not surprisingly the learning rates are essentially the same as in the binary transductive cases, although we note that this analysis was done with Setting 2.

7.1.5 Bounds Based on Learning a Kernel

As a direct extension of the inductive case [Bartlett and Mendelson, 2003], Lanckriet et al. [2004] propose to use the unlabeled data to learn a kernel that is suitable for transductive learning. The idea is to use a kernel method that allows to choose from a certain class of kernels in order to optimize the objective function. The presented PAC-bound shows that good (transductive) performance is achieved with a good trade-off between the complexity of the kernel class and the empirical error. Their exemplary kernel classes are designed as follows. Given an initial set of kernels $\{K_1, ..., K_k\}$, that are defined on the labeled *and* unlabeled data, they define

$$K_c := \{K = \sum_{j=1}^{k} \mu_j K_j \mid K \succcurlyeq 0, \mu_j \in \mathbb{R}, \mathrm{trace}(K) \leq c\}$$

and

$$K_c^{+} := \{K = \sum_{j=1}^{k} \mu_j K_j \mid K \succcurlyeq 0, \mu_j \in \mathbb{R}, \mu_j \geq 0, \mathrm{trace}(K) \leq c\}.$$

Every class of kernels $\mathcal{K}$ give rise to the hypothesis set

$$H_{\mathcal{K}} = \{h(x_j) := \sum_{j=1}^{2n} \alpha_i K_{ij} \mid K \in \mathcal{K}, \alpha = (\alpha_1, ..., \alpha_{2n}) \in \mathbb{R}^{2n}, \alpha^t K \alpha \leq \frac{1}{\gamma^2}\}.$$

The error bound found in this paper reads then as follows.

Theorem 15 (Theorem 24). *For every $\gamma > 0$, with probability at of at least $1 - \delta$ over every training and test set of size n (so $m = n$) uniformly chosen from (X, Y), every function $h \in H_{\mathcal{K}}$ has*

$$R_m(h) \leq \hat{R}_n^{\mathrm{hinge}}(h) + \frac{1}{\sqrt{n}}\left(4 + \sqrt{2\log(\frac{1}{\delta})} + \sqrt{\frac{\mathrm{comp}(\mathcal{K})}{n\gamma^2}}\right),$$

where $\hat{R}^{\mathrm{hinge}}(h)$ is the empirical hinge loss of h and $\mathrm{comp}(\mathcal{K})$ is a complexity measure of $\mathcal{K}$ defined as

$$\mathrm{comp}(\mathcal{K}) = \mathbb{E} \max_{K \in \mathcal{K}} \sigma^t K \sigma$$

with σ being a vector of $2n$ Rademacher variables. The complexity measure for the previously defined kernel classes $\mathcal{K}_c$ and $\mathcal{K}_c^+$ can be computed and bounded by

$$\mathcal{K}_c = c\mathbb{E} \max_{K\in\mathcal{K}} \sigma^t \frac{K}{\operatorname{trace} K}\sigma \leq cn,$$

and

$$\mathcal{K}_c^+ \leq c \min\left(k, n \max_{1\leq j\leq k} \frac{\lambda_j}{\operatorname{trace}(K_j)}\right),$$

where λ_j is the largest eigenvalue of K_j.

Note that since $m = n$ we find that this bound gives the same learning rate of $O(\frac{1}{\sqrt{m+n}})$ as also found in Sections 7.1.3 and 7.1.4.

The effect the unlabeled data has on this procedure depends on the initial kernel guesses $\{K_1, ..., K_k\}$, but is of no further interested in this paper. We can find extensions in Chapelle et al. [2006](p. 282, bottom), where the K_i are chose in a particular way: If we assume that ψ_i is the i-th eigenvector of the graph Laplacian L we can set $K_i = \psi_i {\psi_i}^t$. As described in Chapelle et al. [2006] (p. 280) we can then enforce classifiers found by this procedure to be smooth along the data manifold, if we enforce that μ_i is small when the eigenvalue of ψ_i is large. Similar results are obtained by Johnson and Zhang [2008], where the biggest difference are the kernels that are used. Instead of using an initial set of kernels, Johnson and Zhang [2008] use the spectral decomposition of a given kernel and shrinks it, where the shrinkage depends on the unlabeled data.

7.2 Safe Transductive Learning

In the semi-supervised learning community it is well known that using a semi-supervised procedure often comes with a risk of performance degradation [Chapelle et al., 2006, Chapter 4]. This problem led some authors to ask the question whether it is possible to do semi-supervised learning in a safe way, which means that one can guarantee that the SSL will not be worse than a supervised counterpart. So far we compared mostly SSL and SL risk bounds. But, even if the assumptions of the risk bounds are true, a smaller bound still does not guarantee improvements. We will specifically look at work from Li and Zhou [2011] and Loog [2016]. The results from both works are based on a minimax formulation and show that, under some assumptions, one can indeed guarantee improvements by doing SSL. The analysis is also done in the transductive Setting 1. This means that we have a training set S_n and a test set X_m.

7.2.1 A Minimax Approach for SVMs

The baseline for the model proposed by Li and Zhou [2011] is the $S3VM$ [Bennett and Demiriz, 1999], which takes the unlabeled data into account by finding a low-margin solutions. The proposed model $S4VM$ finds a few diverse low-margin solutions, and then picks amongst these within a minimax framework to hedge against possible worst case scenarios. Assume we found a set of a few proposed solutions $H_p = \{h_1, ..., h_T\}$. The idea is to contrast those solutions to the supervised solutions h^{SVM}. Assume for now that we know the true labels $Y_m = (y_n, ..., y_{n+m})$ of X_m. With this we can calculate the gain and loss in performance when comparing the supervised h^{SVM} to any other classifier h.

$$\operatorname{gain}(h, Y_m, h^{SVM}) := \sum_{i=n}^{n+m} I_{\{h(x_i)=y_i\}} I_{\{h^{SVM}(x_i)\neq y_i\}} \tag{44}$$

$$\operatorname{loss}(h, Y_m, h^{SVM}) := \sum_{i=n}^{n+m} I_{\{h(x_i)\neq y_i\}} I_{\{h^{SVM}(x_i)= y_i\}} \tag{45}$$

If we define our objective as to be the difference of those two

$$J(h, y, h^{SVM}) = \operatorname{gain}(h, Y_m, h^{SVM}) - \operatorname{loss}(h, Y_m, h^{SVM}), \tag{46}$$

we can define a semi-supervised model h^{SSL} as the maximizer of this difference. Since we actually don't know the true labeling, we assume a worst-case scenario that leads to the following max-min formulation.

$$h^{\text{SSL}} = \arg\max_{h\in H_p} \min_{Y\in Y_p} J(h, Y, h^{SVM}) \tag{47}$$

Here $Y_p = \{(h(u_1), ..., h(u_m)) \mid h \in H_p\}$ is the set of all possible labelings that we can achieve with H_p. To guarantee that our SSL is not worse than the SL it is important to assume that the true labels Y_m are part of the set Y_p, because only then we can guarantee the following.

Theorem 16 (Theorem 1)**.** *If $Y_m \in Y_p$, the accuracy of h^{SSL} is never worse than the accuracy of h^{SVM}, when performance is measured on the unlabeled data X_m.*

Again, the crucial assumption is that $Y_m \in Y_p$, which corresponds in this case exactly to a low-density assumption. This is because the set Y_p contains possible labelings that come from classifiers that fulfill the low density assumption. One can imagine to use the same procedure also for different assumptions as we can encode them by Y_p, the set of all labelings that we consider possible. While this paper still needs some assumptions, Loog [2016] shows a case where we get guaranteed improvements assumption-free. This, however, comes at the cost of measuring the improvements in terms of likelihood, and not in terms of accuracy.

7.2.2 A Minimax Approach for Generative Models

The second paper in this line of research is to our knowledge possibly the only paper in semi-supervised learning that considers a completely assumption-free case. This of course comes at a cost, more on that later. The starting point is a family of probability density functions $p(x, y \mid \theta)$ on $\mathcal{X} \times \mathcal{Y}$, where $\theta \in \Theta$ is a parametrization. First we set θ^{SL} to be the supervised maximum likelihood estimator for the model $p(x, y \mid \theta)$, so

$$\theta^{\text{SL}} = \arg\min_{\theta \in \Theta} \left[\sum_{(x,y) \in S_n} \ln p(x, y \mid \theta) \right].$$

Assume for now that we know the true conditional probabilities $p = (p_1, ..., p_{m+n}) \in [0, 1]^{m+n}$ with $p_i = P(Y = 1 \mid X = x_i)$ for $x_i \in S_n \cup X_m$. If we would know this we would actually rather optimize the expected log-likelihood of the model $p(x, y \mid \theta)$ evaluated on the complete dataset $X_{n+m} = \{x_1, ..., x_{n+m}\}$,

$$L(\theta \mid X_{n+m}, p) = \mathbb{E}_{Y \sim p} \left[\sum_{x \in X_{n+m}} \ln p(x, Y \mid \theta) \right]. \tag{48}$$

To be better than the supervised model θ^{sup} on the complete (transductive) likelihood (48) we would like to maximize the likelihood gain over it. So we want to find the θ that maximizes the likelihood gain

$$C(\theta, \theta^{\text{SL}} \mid X_{n+m}, p) = L(\theta \mid X_{n+m}, p) - L(\theta^{\text{SL}} \mid X_{n+m}, p). \tag{49}$$

We cannot maximize (49) directly, since we do not know the class true probability distribution p. We instead set $p(y_i \mid x_i) = 1$ for all labeled points $(x_i, y_i) \in S_n$ which gives us the vector $p_n = (p(1 \mid x_1), ..., p(1 \mid x_n))$ and for the unlabeled points X_m we consider a worst case, which leads to the following max-min formulation.

$$\theta^{\text{SSL}} = \arg\max_{\theta \in \Theta} \min_{p_m \in [0,1]^m} C(\theta, \theta^{\text{SL}} \mid X_{n+m}, (p_n, p_m)) \tag{50}$$

Note that the vector p_m can be the true labels Y_m of the unlabeled data X_m. Note also that $C(\theta^{\text{SSL}}, \theta^{\text{SL}} \mid X_{n+m}, (p_n, p_m) \geq 0$ for all $p_m \in [0, 1]^m$, so in particular if $p_m = Y_m$, as we can always chose $\theta^{SSL} = \theta^{\text{SL}}$. That means that the following theorem holds.

Theorem 17 (Lemma 1)**.** *Let θ^{SSL} be a solution found in Equation* (50)*, then*

$$L(\theta^{\text{SL}} \mid X_{n+m}, Y_{n+m}) \leq L(\theta^{\text{SSL}} \mid X_{n+m}, Y_{n+m}), \tag{51}$$

and for some specific choices for the model $p(x, y \mid \theta)$ the previous inequality is almost surely strict. So we are guaranteed that the transductive likelihood of our semi-supervised model is larger than of the supervised model.

An important difference between this work and the previous section is that for this paper one employs a generative model $p(x, y)$, while the SVM used by Li and Zhou [2011] is a discriminative model that inherently optimizes the class probability $p(y \mid x)$. Krijthe and Loog [2018], see also Subsection 3.1.4, show that to some degree it is actually necessary to use a generative model: The semi-supervised estimator of Equation (50) will coincide with the supervised estimator for a large class of discriminative models. There are several explanations why a joint model $p(x, y)$ helps out in the situation. The intuitive and obvious one is that the likelihood of this model takes the marginal distribution $P(X)$ into account, a quantity that can be measured from unlabeled data.

8 Discussion

We covered the main theoretical ideas and results that have been put forward over the past four decades in the field of semi-supervised learning. Specifically, we focused on results that inform us about its potential and the lack of such potential. We covered the answers to the questions: What are the limits of semi-supervised learning? What are the assumptions of different methods? What can we achieve if the assumptions are true? We like to wrap up our survey and mention a few realizations that, we think, get to the core of it.

8.1 On The Limits of Assumption Free SSL

In Section 3 we reviewed work that analyzes the limits of semi-supervised learning when no particular assumptions about the distribution are made, which a semi-supervised learner can exploit. The most general formulation of this is captured in Conjecture 1 and 2. They essentially state that a semi-supervised learner can beat all supervised learners by at most a constant. We then presented work that shows that the conjectures do not hold in full generality, but in particular situations. They essentially hold for the realizable case and hypothesis classes of finite VC-dimension, while they do not hold in the realizable or agnostic case for infinite VC-dimension. It remains to investigate the case of agnostic PAC-learning with a finite VC-dimension.

8.2 How Good Can Constant Improvement Be?

The question studied in Section 3.1.6 and the previous Subsection is whether a semi-supervised learner can offer more than a constant improvement, in terms of sample complexity. One can, however, also ask the question how good already a constant improvement can be in practice. The answer to that can be seen through a thought experiment. Assume that we have two classes given by two concentric d-dimensional spheres. Assume that we have enough unlabeled data for a manifold regularization scheme to identify the spheres. With this the semi-supervised learner needs only one labeled sample per class to give a perfect classification, while every supervised learner needs for good generalization a labeled sample size which increases in the dimension d. Although the manifold regularized classification needs only two samples, we know from Mey et al. [2019] that manifold regularization can only achieve constant improvement. This might seem contradictory, but this behavior is easily understood when we study the VC-dimensions. If the supervised classifier uses a hypothesis space H, we can interpret manifold regularization as switching to a restricted space $\tilde{H}_\lambda$. This space only contains hypotheses that fulfill a manifold assumption, where the regularization parameter λ indicates to which degree this assumption is enforced. Mey et al. [2019] show that the improvement of using manifold regularization is at most $\mathrm{VC}(H)/\,\mathrm{VC}(\tilde{H}_\lambda)$. If we set λ high enough we can keep $\mathrm{VC}(\tilde{H}_\lambda)$ constant, while $\mathrm{VC}(H)$ will increase with the dimension d. This shows that the constant improvement can be arbitrarily high. While this example uses the manifold assumption, Golovnev et al. [2019] give a example with a semi-supervised learner that has the full knowledge of the domain distribution. We explain the particular example in Section 3.1.6. This shows that the constant improvement can be arbitrarily high if we have further assumptions, like the manifold assumption, or full knowledge of the marginal distribution. It is an open question if one can have arbitrarily high constants without assumptions and with limited unlabeled data.

8.3 The Amount of Unlabeled Data We Need

In Section 3.2 we presented three settings, in which a semi-supervised learner can PAC-learn, while no supervised learner can. For that we need, in principle, an infinite amount of unlabeled data and we also cannot create an example where that is not the case. If a fixed finite amount of unlabeled data would be enough to learn under any given distribution P we could just use the same strategy to learn in a supervised way as we can always chose to ignore the label. The way those examples work is that for each fixed P a finite amount of unlabeled data is sufficient, but this amount can be arbitrarily large. This has the consequence that if we want to learn over all possible distributions we need an arbitrarily large amount ($= \infty$) of unlabeled data. The improvements that semi-supervised learning can offer which we present in Sections 4, 5 and 6 do not necessarily need an infinite amount of unlabeled data, although it sometimes assumed for convenience. The difference is that in those settings supervised learner are also able to PAC-learn, but a semi-supervised learner is able to do this with fewer labeled samples. In Sections 6.2 and 6.3 we saw two instantiations of a cluster assumption, and the authors showed that the amount of unlabeled data needs to increase exponentially with the amount of labeled data to make use of this assumption. This is because the error in finding the clusters decreases only polynomially in the number of unlabeled points as shown in Inequality 31.

8.4 Using Assumptions in Semi-Supervised Learning

In Sections 5 and 6 we investigate what a semi-supervised learner can achieve once assumptions are made. A semi-supervised assumption is a link between the domain distribution and the labeling function. In particular we assume that we can ignore certain labeling functions after we have seen a specific domain distribution. The cluster assumption, for example, would exclude labeling functions that do not assign the same label to points belonging to the same cluster. The obvious, but real problem with this is that we do not know if such assumptions do hold or not. We speculate that testing if such an assumption is true or not consumes as many labeled points as learning directly a good classification rule with a supervised learner. To make this statement precise we define an assumption as a property of the distribution P on $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{P}^A$ be a set of distributions on $\mathcal{X} \times \mathcal{Y}$. We say that P fulfills assumption A iff $P \in \mathcal{P}^A$. For example $\mathcal{P}^A$ could only contain distributions such that the marginal distributions $P(X)$ have always support on clusters, and

symbol	explanation
$\mathcal{X}$	Feature space, for example $\mathcal{X} = \mathbb{R}^n$
$\mathcal{Y}$	Label space. Classification: $\mathcal{Y} = \{-1, 1\}$. Regression: $\mathcal{Y} = \mathbb{R}$.
P	Distribution on $\mathcal{X} \times \mathcal{Y}$
$\mathcal{P}$	A set of distributions on $\mathcal{X} \times \mathcal{Y}$
X, Y	Random variables distributed according to P
$P(X)$	Marginal distribution of P w.r.t to $\mathcal{X}$
$P(Y)$	Marginal distribution of P w.r.t to $\mathcal{Y}$
D	Domain distribution on $\mathcal{X}$
$\mathcal{D}$	A set of domain distributions on $\mathcal{X}$
$I_{\{\text{Boolean expression}\}}$	Indicator function (equals 1 if expression is true and 0 else)
$l(\hat{y}, y)$	Loss function, if not specified otherwise $l(\hat{y}, y) = I_{\hat{y}=y}$
H	Hypothesis class, where each $h \in H$ is a map $h : \mathcal{X} \to \mathcal{Y}$
$R(h)$	The risk of $h \in H$. Precisely: $R(h) = \mathbb{E}_{X,Y}[l(h(X), y)]$
(x_i, y_i)	A realization of (X, Y)
S_n	A labeled sample set of size n, $S_n = ((x_1, y_1), ..., (x_n, y_n))$
U_m	A unlabeled sample set of size m, usually $U_m = \{x_{n+1}, ..., x_{n+m}\}$
$\hat{R}_n(h) = \hat{R}(h)$	Empirical risk of h w.r.t S_n, $\hat{R}(h) = \frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)$
h^{SSL}	Model trained on S_n and U_m or $P(X)$, where $h^{\text{SSL}} : \mathcal{X} \to \mathcal{Y}$
h^{SL}	Model trained on S_n, where $h^{\text{SL}} : \mathcal{X} \to \mathcal{Y}$
$m(H, \epsilon, \delta)$	Supervised sample complexity, see Definition 1
$m^{\text{SSL}}(H, \epsilon, \delta)$	Semi-supervised sample complexity, see Definition 2

Table 1: Complete list of notations used in this survey.

each cluster has a unique label. Then P fulfills this particular cluster assumption A iff $P \in \mathcal{P}^A$. The crucial thing to note is, that the assumption A is a property on P, so we need labeled samples to test whether its true or not. It is thus of interest to compare the consumption of labeled data for reducing the uncertainty about the assumption to the consumption of labeled data for the convergence of the semi-supervised learner. We might of course know a priori that the assumption is true and do not need to test it, but what if not?

One of the few works that analyze this is reviewed in Section 6.4. Azizyan et al. [2012] show that one can get essentially faster rates if the assumption is true, but we pay a penalty of $O(\frac{\ln(n)}{n})$ if it is not true. Balcan et al. [2011] investigates how one can test for a property in an active way, so when we can choose which samples we want to label. The implications of this testing procedure for semi-supervised learning are, however, not clear. Of course, we may claim that it is not even necessary to test if the assumption is true or not, following Vapnik's principle: Why should we test if the assumption is true or not, when we are ultimately only interested whether the semi-supervised learner performs better or not? We believe that this is an important open question in semi-supervised learning.

9 Definitions

Definition 1. ***Supervised Sample Complexity*** *Given a learning problem (P, l, H) and $\epsilon, \delta > 0$ we define the sample complexity $m(B, H, P, \epsilon, \delta) \in \mathbb{N}$ of a supervised learner B as the smallest natural number k such that with probability at least $1 - \delta$ over all possible draws of a labeled sample S_k it holds that*

$$R(B(S_k)) - \inf_{h \in H} R(h) \leq \epsilon.$$

Or in short

$$m(B, H, P, \epsilon, \delta) = \{\min k \in \mathbb{N} \mid P\left(R(B(S_k)) - \inf_{h \in H} R(h) \leq \epsilon\right) \geq 1 - \delta\}.$$

Although not explicitly mentioned in the definition above, if B is semi-supervised it has additional input in form of either $P(X)$, or a random draw from it. Sometimes we drop the learner B from the sample complexity notation $m(B, H, P, \epsilon, \delta)$, and write either $m(H, P, \epsilon, \delta)$ or $m^{\text{SSL}}(H, P, \epsilon, \delta)$ if there exists a supervised or semi-supervised learner respectively that achieves the sample complexity.

Definition 2. ***Semi-Supervised Sample Complexity*** *Given a learning problem (P, l, H) and $\epsilon, \delta > 0$ we define the sample complexity $m^{\text{SSL}}(B, H, P, \epsilon, \delta) \in \mathbb{N}$ of a semi-supervised learner B, which has information about the marginal in the form of $U \in \{U_m, P(X)\}$, as the smallest natural number k such that with probability at least $1 - \delta$ over all possible draws of a labeled sample S_k it holds that*

$$R(B(S_k, U)) - \inf_{h \in H} R(h) \leq \epsilon.$$

Or in short

$$m^{\text{SSL}}(B, H, P, \epsilon, \delta) = \{\min k \in \mathbb{N} \mid P\left(R(B(S_k), U) - \inf_{h \in H} R(h) \leq \epsilon)\right) \geq 1 - \delta\}.$$

We usually drop the learner B from the sample complexity notation $m(B, H, P, \epsilon, \delta)$, and write either $m(H, P, \epsilon, \delta)$ or $m^{\text{SSL}}(H, P, \epsilon, \delta)$ if there exists respectively a supervised or semi-supervised learner that achieves this sample complexity. Similarly we drop the distribution P from the notation and write $m(H, \epsilon, \delta)$ or $m^{\text{SSL}}(H, \epsilon, \delta)$ if we can achieve this sample complexity for all distributions P.

References

Martin Azizyan, Aarti Singh, and Larry A. Wasserman. Density-sensitive semisupervised inference. *Computing Research Repository*, abs/1204.1685, 2012.

Maria-Florina Balcan and Avrim Blum. A discriminative model for semi-supervised learning. *Journal of the ACM*, 57 (3):19:1–19:46, 2010.

Maria-Florina Balcan, Eric Blais, Avrim Blum, and Liu Yang. Active testing. *Computing Research Repository*, abs/1111.0897, 2011.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, March 2003. ISSN 1532-4435.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 08 2005.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, 7:2399–2434, December 2006. ISSN 1532-4435.

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 The 21st Annual Conference on Learning Theory*, Helsinki, Finland, 2008.

Gyora M. Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2): 377–389, September 1991. ISSN 0304-3975.

Kristin P. Bennett and Ayhan Demiriz. Semi-supervised support vector machines. In *Advances in Neural Information Processing Systems 11*, pages 368–374, Denver, CO, USA, 1999.

Peter J. Bickel and Bo Li. *Local polynomial regression on unknown manifolds*, volume Volume 54, pages 177–186. Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2007.

Avrim Blum and Shuchi Chawla. Learning from labeled and unlabeled data using graph mincuts. In *Proceedings of the 18th International Conference on Machine Learning*, pages 19–26, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc.

Avrim Blum and Tom Mitchell. Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th Annual Conference on Computational Learning Theory*, pages 92–100, Madison, Wisconsin, USA, 1998.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499–526, March 2002. ISSN 1532-4435.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, New York, NY, USA, 2004.

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

Vittorio Castelli and Thomas M. Cover. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. *IEEE Transactions on Information Theory*, 42(6):2102–2117, 1996.

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. The MIT Press, Cambridge, MA, USA, 2006.

Fabio Cozman and Ira Cohen. Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers. In *Semi-Supervised Learning*, chapter 4, pages 57–72. The MIT Press, Cambridge, MA, USA, 2006.

Malte Darnstädt, Hans Ulrich Simon, and Balázs Szörényi. Unlabeled data does provably help. In *Symposium on Theoretical Aspects of Computer Science*, volume 20, pages 185–196, Kiel, Germany, 2013.

Philip Derbeko, Ran El-Yaniv, and Ron Meir. Explicit learning curves for transduction and application to clustering and compression algorithms. *Computing Research Repository*, abs/1107.0046, 2011.

Ran El-Yaniv and Dmitry Pechyony. Stable transductive learning. In *19th Annual Conference on Learning Theory*, volume 4005, pages 35–49, Pittsburgh, PA, USA, 2006.

Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35(1):193–234, June 2009. ISSN 1076-9757.

Jason Farquhar, David Hardoon, Hongying Meng, John S. Shawe-taylor, and Sándor Szedmák. Two view learning: Svm-2k, theory and practice. In *Advances in Neural Information Processing Systems 18*, pages 355–362. MIT Press, 2006.

Amir Globerson, Roi Livni, and Shai Shalev-Shwartz. Effective semisupervised learning on manifolds. In *Conference on Learning Theory 2018*, pages 978–1003, Amsterdam, The Netherlands, 2017.

Alexander Golovnev, Dávid Pál, and Balázs Szörényi. The information-theoretic value of unlabeled data in semi-supervised learning, 2019.

Christina Göpfert, Shai Ben-David, Olivier Bousquet, Sylvain Gelly, Ilya Tolstikhin, and Ruth Urner. When can unlabeled data improve the learning rate?, 2019.

Lars Kai Hansen. *On Bayesian Transduction: Implications for the Covariate Shift Problem*, page 65–72. The MIT Press, Cambridge, MA, USA, 2009.

Thorsten Joachims. Transductive inference for text classification using support vector machines. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 200–209, San Francisco, CA, USA, 1999.

R. Johnson and Tong Zhang. Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, 54(1):275–288, January 2008. ISSN 0018-9448.

Rie Johnson and Tong Zhang. On the effectiveness of laplacian normalization for graph semi-supervised learning. *Journal of Machine Learning Research*, 8:1489–1517, December 2007. ISSN 1532-4435.

Matti Kääriäinen. Generalization error bounds using unlabeled data. In *18th Annual Conference on Learning Theory*, pages 127–142, Bertinoro, Italy, 2005. Springer.

Masanori Kawakita and Takafumi Kanamori. Semi-supervised learning with density-ratio estimation. *Machine Learning*, 91(2):189–209, 2013.

Jesse Krijthe and Marco Loog. The pessimistic limits of margin-based losses in semi-supervised learning. In *Advances in Neural Information Processing Systems 31*, Montreal, Canada, 2018.

Julius Von Kügelgen, Alexander Mey, and Marco Loog. Semi-generative modelling: Covariate-shift adaptation with cause and effect features. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1361–1369, Okinawa, Japan, 2019.

Samuel Kutin. Extensions to mcdiarmid's inequality when differences are bounded with high probability. Technical report, University of Chicago, 2002.

Samuel Kutin and Partha Niyogi. Almost-everywhere algorithmic stability and generalization error. *Computing Research Repository*, abs/1301.0579, 2013.

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

Gert R. G. Lanckriet, Nello Cristianini, Peter L. Bartlett, Laurent El Ghaoui, and Michael I. Jordan. Learning the kernel matrix with semidefinite programming. *Journal of Machine Learning Research*, 5:27–72, 2004.

John Langford and John Shawe-Taylor. Pac-bayes & margins. In *Advances in Neural Information Processing Systems 15*, pages 439–446, Vancouver, British Columbia, Canada, 2002.

B. Leskes. The value of agreement, a new boosting algorithm. In *Proceedings of the 18th Conference on Learning Theory*, Bertinoro, Italy, 2005.

Yu-Feng Li and Zhi-Hua Zhou. Towards making unlabeled data never hurt. In *Proceedings of the 28th International Conference on Machine Learning*, pages 1081–1088, Bellevue, Washington, USA, 2011.

Marco Loog. Contrastive pessimistic likelihood estimation for semi-supervised classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):462–475, 2016.

Yury Maximov, Massih-Reza Amini, and Zaïd Harchaoui. Rademacher complexity bounds for a penalized multiclass semi-supervised algorithm. *Computing Research Repository*, abs/1607.00567, 2016.

David McAllester. Simplified pac-bayesian margin bounds. In *Learning Theory and Kernel Machines*, pages 203–215, Berlin, Heidelberg, 2003a. Springer Berlin Heidelberg.

David A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, Apr 2003b. ISSN 1573-0565.

David A. McAllester. Pac-bayesian stochastic model selection. *Machine Learning*, 51(1):5–21, Apr 2003c. ISSN 1573-0565.

Alexander Mey, Tom Viering, and Marco Loog. A distribution dependent and independent complexity analysis of manifold regularization, 2019.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. The MIT Press, Cambridge, MA, USA, 2012.

Partha Niyogi. Manifold regularization and semi-supervised learning: some theoretical analyses. *Journal of Machine Learning Research*, 14(1):1229–1250, 2013.

Dmitry Pechyony. *Theory and Practice of Transductive Learning*. PhD thesis, Isreal Institute of Technology, 2008.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. The MIT Press, Cambridge, MA, USA, 2017.

Joel Ratsaby and Santosh S. Venkatesh. Learning from a mixture of labeled and unlabeled examples with parametric side information. In *Proceedings of the 8th Annual Conference on Computational Learning Theory*, pages 412–417, Santa Cruz, CA, USA, 1995.

Philippe Rigollet. Generalization error bounds in semi-supervised classification under the cluster assumption. *Journal of Machine Learning Research*, 8:1369–1392, 2007.

David S. Rosenberg and Peter L. Bartlett. The rademacher complexity of co-regularized kernel classes. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 396–403, San Juan, Puerto Rico, 21–24 Mar 2007.

B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1255–1262, New York, NY, USA, 2012. Omnipress.

Matthias Seeger. Input-dependent Regularization of Conditional Density Models. Technical report, Institute for Adaptive and Neural Computation, 2000.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014.

Vikas Sindhwani and David S. Rosenberg. An rkhs for multi-view learning and manifold co-regularization. In *Proceedings of the 25th International Conference on Machine Learning*, pages 976–983, Helsinki, Finland, 2008.

Aarti Singh, Robert D. Nowak, and Xiaojin Zhu. Unlabeled data: Now it helps, now it doesn't. In *Advances in Neural Information Processing Systems 21*, pages 1513–1520, Vancouver, British Columbia, Canada, 2008.

Kaushik Sinha and Mikhail Belkin. The value of labeled and unlabeled examples when the model is imperfect. In *Advances in Neural Information Processing Systems 20*, pages 1361–1368, Vancouver, British Columbia, Canada, 2007.

Nataliya Sokolovska, Olivier Cappé, and François Yvon. The asymptotics of semi-supervised learning in discriminative probabilistic models. In *Proceedings of the 25th International Conference on Machine Learning*, volume 307, pages 984–991, Helsinki, Finland, 2008.

Karthik Sridharan and Sham M. Kakade. An information theoretic framework for multi-view learning. In *21st Annual Conference on Learning Theory*, pages 403–414, Helsinki, Finland, 2008.

Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1): 135–166, 02 2004.

Vladimir Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, Heidelberg, 1982.

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

Tong Zhang and Frank J. Oles. A probability analysis on the value of unlabeled data for classification problems. In *Proceedings of the 17th International Conference on Machine Learning*, Stanford, CA, USA, 2000.