Improved Generalization in Semi-Supervised Learning
A Survey of Theoretical Results

Mey, Alexander; Loog, Marco

DOI
10.1109/TPAMI.2022.3198175

Publication date
2022

Document Version
Final published version

Published in
IEEE Transactions on Pattern Analysis and Machine Intelligence

Improved Generalization in Semi-Supervised Learning: A Survey of Theoretical Results

Alexander Mey and Marco Loog

Abstract—Semi-supervised learning is the learning setting in which we have both labeled and unlabeled data at our disposal. This survey covers theoretical results for this setting and maps out the benefits of 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, including unlabeled data may actually decrease performance. For all practical purposes, it is therefore instructive to have an understanding of the underlying theory and the possible learning behavior that comes with it. This survey gathers results about the possible gains one can achieve when using semi-supervised learning as well as results about the limits of such methods. Specifically, it aims to answer 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? As, indeed, the precise assumptions made are of the essence, this is where the survey's particular attention goes out to.

Index Terms—Semi-supervised learning, learning theory, improvement guarantees, assumptions

1 Introduction

For many applications, gathering unlabeled data is faster and cheaper than gathering labeled data. The goal of semi-supervised learning (SSL) is to combine both and design classification and regression rules that outperform schemes only based on labeled data. SSL does come, however, with an inherent risk: including unlabeled data can also degrade performance [1], [2]. Studying and understanding SSL from a theoretical point of view allows one to formulate the necessary assumptions, the expected improvements, and the limitations of the different methods. Based on such understanding, one can formulate recommendations for using SSL with the aim of avoiding any decrease in performance as good as possible. Our review provides this theoretical viewpoint, offering a much-needed complement to claims that there are no performance guarantees (see, for instance, [3, page 380]). We study the relevant, theoretical papers in detail, present their main findings, and point out connections. Next to theoretical guarantees of some specific learners, we also cover the theoretical limits of SSL.

- *Alexander Mey is with the Delft University of Technology, 2628 Delft, Netherlands. E-mail: a.mey@tudelft.nl.*
- *Marco Loog is with the Delft University of Technology, 2628 Delft, Netherlands, and also with the University of Copenhagen, 1165 Copenhagen, Denmark. E-mail: m.loog@tudelft.nl.*

Manuscript received 12 February 2021; revised 2 June 2022; accepted 28 July 2022. Date of publication 15 August 2022; date of current version 6 March 2023.
This work was supported by the Netherlands Organisation for Scientific Research TOP under Grant 612.001.402.
(Corresponding author: Alexander Mey.)
Recommended for acceptance by S. Kaski.
This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2022.3198175, provided by the authors.
Digital Object Identifier no. 10.1109/TPAMI.2022.3198175

1.1 Common Assumptions

Much in this survey revolves around making precise what assumptions underlie which results. Foregoing such precision for now, this subsection introduces the most common ones and sketches their relation. Conceptually, most assumptions restrict how the data may be labeled, given a specific domain distribution. This concept will often reappear in this survey, Section 2.1.6 in particular investigates the effectiveness of SSL if such assumptions are not made.

One of most used assumptions is the *smoothness assumption* [4, Section 1.2]. It roughly states that two input points that are close together, have a high likelihood to share the same output. The important word is *close*. One could call two points close, when their Euclidean distance is small, but one can think of more sophisticated ways to define closeness. One way is through the *cluster assumption*. The idea is that we can use the unlabeled data to find clusters and call two points close if they are in the same cluster. Section 5 formalizes this and shows the assumption to be very strong, i.e., it enables exponentially fast learning.

Low-density separation can be seen as a specific instance of the cluster assumption, but giving rise to different algorithms. It states that the decision boundary should lie in a region with low density. Indeed, if we define clusters as regions of high density and would like to separate those, the decision boundary should automatically be in a low-density region. Again, the unlabeled data helps, as we can actually identify the low-density regions as, for example, formalized through the transductive support vector machine [5], [6].

The *manifold assumption* is related to the above concepts, but has led to confusion as there are two alternative definitions. The first is best explained with a quote from [7]: "We will assume that if two points $x_1, x_2 \in X$ are close in the intrinsic geometry of $P(X)$, then the conditional distributions $P(y|x_1)$ and $P(y|x_2)$ are similar." The manifold refers to

this intrinsic geometry of $P(X)$. Importantly, note that this is the same as the cluster assumption, the cluster is formalized as the manifold geometry given by $P(X)$. An alternative definition has, for example, been given in [4, Section 1.2.3]: "The (high-dimensional) data lie (roughly) on a low dimensional manifold." Note that this definition does not *not* restrict how one may label the data, given the domain distribution. Although a low-dimensional manifold can help to avoid the curse of dimensionality, Section 2.1.5 reveals that such knowledge does not bring any additional advantage regarding worst-case performance rates, also called minimax rates. If not stated otherwise, the first definition is used.

There are but few assumptions that really diverge from the above concepts. A notable exception is the multi-view assumption, which essentially states that one can split the feature space into two subspaces with each subspace being sufficient to solve the learning problem. Section 4.2 covers one formalization and explains the intuition of how this assumption can help the learning process.

What all of those assumptions have in common is that it is unclear if we can effectively verify them, or if for successful semi-supervised learning they have to be known to hold in advance, see also Section 7.4.

1.2 Outline

Section 2 discusses results on the limits of SSL, which typically arise form specific assumptions about the model or the data generation process. As opposed to provably limited improvements, this same section presents three settings where the improvements of SSL are *unlimited*, i.e., where a semi-supervised learner can learn the problem, while no supervised learner (SL) can. Section 3 investigates what is possible with some specific methods that exploit unlabeled data without making further assumptions on the data distribution. Section 4 treats semi-supervised learners that make *weak* assumptions on the data distribution, in the sense that the resulting learner cannot get a learning rate faster than $\frac{1}{\sqrt{n}}$, with n the number of labeled samples.[1] Here, improvements are given by a constant. Section 5 then discusses learners that use *strong* assumptions, providing converge exponentially fast to the best classifier in a given class, i.e., the learning rate is of the order e^{-n}. This section also argues that there is not necessarily a principled qualitative difference in weak and strong assumptions, but rather a subtle quantitative difference. Subsequently, Section 6 presents results in the transductive setting where one is only interested in the labels of the unlabeled data available. The same section present a line of research that aims to construct semi-supervised learners that are never worse than their supervised counterparts. Finally, Section 7 discusses the overall results and conclude with what we see as the current challenges in the field. Next to that, it reconsiders what it means to use assumptions and the problems that come with it. This final section also makes note of the absence of deep learning from this review. Before turning to Section 2, the next subsection briefly introduces the formal learning framework that is assumed in most of the remainder.

1.3 The Learning Framework

We typically present results, describing the performance of semi-supervised learners, in the language of PAC-learning.[2] Unless specified otherwise, we consider a standard statistical learning setting: we are given a feature space $\mathcal{X}$ and an output space $\mathcal{Y}$, together with an unknown distribution P on $\mathcal{X} \times \mathcal{Y}$. With slight abuse of notation, we write $P(X)$ and $P(Y)$ for the marginal distributions on $\mathcal{X}$ and $\mathcal{Y}$. Similar conventions are used for conditional distributions.

We consider the setting in which we have observed a labeled n-sample $S_n = ((x_1, y_1), \ldots, (x_n, y_n))$ and an unlabeled m-sample $U_m = (x_{n+1}, \ldots, x_{n+m})$, 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 chooses a hypothesis class H, where each $h \in H$ is a mapping $h : \mathcal{X} \to \mathcal{Y}$, and a loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. A (semi)-supervised learner B is a map that receives as input the labeled (and unlabeled) sample S_n, U_m and maps to hypothesis h, so $B(S_n, U_m) \in \mathcal{H}$. A strictly supervised learner receives an empty second input. Unless specified otherwise, we assume for classification that $\mathcal{Y} = \{-1, +1\}$ and the loss is the 0-1 loss: $l(y, \hat{y}) = I_{\{y \neq \hat{y}\}}$. For the regression task, we assume that $\mathcal{Y} = \mathbb{R}$ and consider the standard squared loss: $l(y, \hat{y}) = (y - \hat{y})^2$. Based on the n labeled and m unlabeled samples, the aim is to find an $h \in H$ such that the risk $R(h) := \mathbb{E}_{X,Y}[l(h(X), Y)]$ is small.

Whenever we have any quantity A that depends on the distribution P, we write $\hat{A}$ for an 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. It should be clear from the context on which sample we measure the loss.

Finally we denote by $m(\cdots)$ and $m^{\mathrm{SSL}}(\cdots)$ the supervised and semi-supervised sample complexity, as defined in the appendix, which can be found on the Computer Society Digital Library at http://doi.ieeecomputersociety.org/10.1109/TPAMI.2022.3198175.

2 Possibilities and Impossibilities

In SSL, we want to use information about the distribution on $\mathcal{X}$ to improve learning. It is not directly clear, however, that this information is useful at all. Various works formalize the idea of using unlabeled data and subsequently investigate situations where unlabeled data cannot help or where it, in fact, can. This section follows the same division between impossibility (Section 2.1) and possibility (Section 2.2). The latter presents three specific settings where unlabeled data can, in fact, give *unlimited* improvement, i.e, no supervised learner can PAC-learn in the situation considered, whereas some semi-supervised learner can.

1. The learning rate is the rate at which a learner converges to the best classifier in a given class. Without further assumptions, the standard rate of the order $\frac{1}{\sqrt{n}}$ follows from classic learning results [5], [8], [9].

2. PAC-learning stands for *Probably Approximately Correct*-learning. This framework studies how far a trained classifier is off from the best classifier in a class given a certain amount of labeled data. Good introductions to this framework can be found in [8] and [9]. For completeness, Definition 1 introduces the notion of sample complexity. PAC-learnable means that the sample complexity is always finite.

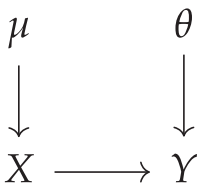


Fig. 1. Data generation process used in [11].

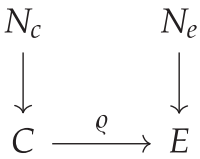


Fig. 2. Simple functional causal model [16]. The effect E is caused by C given a deterministic mapping ϱ. E and C are influenced by noise variables N_E and N_C, respectively.

We note that the negative results often assert an independence between the posterior probability $P(Y|X)$ 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 $P(Y|X)$ but in the complete risk $\mathbb{E}_{X,Y}[l(h(X), Y)]$ of a classifier h, which *does* depend on $P(X)$ [10, Subsection 5.1.2]. Sections 3.1 and 3.2, for example, present works that show risk improvements even when $P(Y|X)$ and $P(X)$ are independent.

2.1 Impossibility Results

The results covered in this subsection show, in different settings, that semi-supervised learning is inherently impossible. While the titles in the following section indicate the setting that renders semi-supervised learning impossible, we often reference later sections that explicitly exclude this setting to generate positive results. Next to this, this section presents results that demonstrate the limits of semi-supervised learning methods when no particular assumptions about the data distribution are made (cf. Section 1.1).

2.1.1 Due to Data Generation Process

Ref. [11] 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. Firstly, the distribution P comes from a model class with parameters μ and θ. Subsequently, values $\mu \sim P_\mu$ and $\theta \sim P_\theta$ are sampled independently after which the data is generated by gathering samples $x \sim P(X|\mu)$ with corresponding labels $y \sim P(Y|X, \theta)$, see also Fig. 1.

The goal is to infer θ from a finite labeled sample $S_n = (x_i, y_i)_{1 \leq i \leq n}$. It can be easily shown that $P(\theta|S_n)$ 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 is to assume a dependency between μ and θ. This exact approach was chosen in [12, Example 1] to create a setting where knowledge of the marginal distribution can indeed help. In their example, the marginal distribution completely determines the Bayes classifier. Therefore, a semi-supervised learner exists that always has zero risk, while any supervised learner has the standard learning rate of $\frac{1}{\sqrt{n}}$. Alternatively, we can also think about settings where the data generation process from Fig. 1 is reversed: first sample a label y, and then sample a feature x from a marginal distribution associated to y, a setup we cover in Section 5.1.

2.1.2 Due to Model Assumptions

Ref. [13] investigates when unlabeled data should change our posterior belief about a model. In comparison to [11], no data generation assumptions are made, but rather assumptions about the model that is used. The author looks at solutions derived from the expected squared loss between this given model and the true desired label output. Splitting the joint distribution $P(X, Y|\theta)$ of the model considered as $P(X, Y|\theta) = P(Y|X, \theta_1, \theta_2)P(X|\theta_2, \theta_3)$, the conclusion is reached that unlabeled data can be discarded if θ_2, the shared parameter between the label and marginal distribution, is empty.

Conversely, the effectiveness of methods like expectation maximization [14] or the provable improvements of the method from Section 6.2.2 stem from the fact that some generative models *cannot* be decomposed in the above way. Given, for example, data that is distributed as two Gaussian distributions, where each distribution corresponds to a class. This means that $\theta = \{q, \mu_1, \mu_2, \Sigma_1, \Sigma_2\}$, with μ_i and Σ_i ($i \in \{1, 2\}$) the class means and covariance matrices, and $q \in [0, 1]$ the class prior of, say, class 1. Here both $P(Y|X, \theta)$ and $P(X|\theta)$ depend on the class means and covariances.

Earlier work, [15], distinguishes the same type of models, but the impossibility is about the asymptotic efficiency of semi-supervised classifiers. Specifically, it considers the following two joint probabilities, which both provide generative models: parametric: $P(X, Y|\alpha) = P(X|\alpha)P(Y|X, \alpha)$; semi-parametric: $P(X, Y|\alpha) = P(X)P(Y|X, \alpha)$. The author shows that the Fisher information $I(\hat{\alpha})_{\text{unlab} + \text{lab}}$ of an maximum likelihood estimator (MLE) $\hat{\alpha}$ that takes labeled and unlabeled data into account can be decomposed as $I(\hat{\alpha})_{\text{unlab} + \text{lab}} = I(\hat{\alpha})_{\text{unlab}} + I(\hat{\alpha})_{\text{lab}}$. So, as long as unlabeled data is available, the Fisher information of the semi-supervised learner is larger compared to the supervised learner, as the latter equals $I(\hat{\alpha})_{\text{lab}}$. It follows that the semi-supervised learner is asymptotically more efficient, although not necessarily strictly. In the parametric case, $I(\hat{\alpha})_{\text{unlab}} = 0$ and the semi-supervised and supervised estimator have the same asymptotic behavior. The primary difference to the previous subsection is that now we have an impossibility of gain in Fisher information, rather than one of Bayes updating.

2.1.3 Because of Causal Direction

Ref. [16] analyzes a functional causal model, such as the one in Fig. 2. Different learning scenarios are considered under the assumption that the label is the cause C and the feature is the effect E and vice versa. This model introduces an asymmetry in cause and effect, since it leads to the fact that $P(C)$ and $P(E|C)$ are independent, while $P(E)$ and $P(C|E)$ are not. Assuming that X is the cause of the label Y, the prediction $P(Y|X)$ is independent of newly gained information

about $P(X)$. This independence vanishes, if we assume that the label Y is caused by X. This excludes the possibility to improve the posterior prediction $P(Y|X)$ with the help of unlabeled data. However, as mentioned in the beginning of Section 2, the unlabeled data may still help to reduce the risk, as the risk always depends on $P(X)$.

2.1.4 To Always Outperform Supervised Learner

Inspired by a successful approach for a *generative* linear discriminant model from [17] (see Section 6.2.2), [18] investigates a similar approach to find semi-supervised solutions for *discriminative* models that are never worse than their supervised counterparts. Discriminative models are considered that use a monotonously decreasing loss function, while the setting is transductive, i.e., interest is in the performance of the model on the unlabeled data U_m only. (Section 6 discusses this setting in more detail.) The work essentially shows that, under some mild conditions, there is always a labeling of the unseen data U_m such that a semi-supervised learner performs worse on U_m than the supervised solution does. It is impossible, therefore, to guarantee that the semi-supervised solution always outperforms the supervised solution.

2.1.5 Only Knowing the Manifold

Ref. [19] shows that knowledge of the manifold alone, without additional assumption, is not sufficient to outperform a purely supervised learner (cf. Section 1.1 the second definition of manifold assumption). It works in a regression setting and extends work in [20], which introduces a supervised learner that performs regression on an unknown manifold, to show that there is a supervised learner that can adapt to the dimension of the manifold and thus can achieve worst case rates, also called minimax rates, equivalent to a learner that directly works on the lower dimensional manifold.

We note that [19] also shows that one can achieve essentially faster rates by making a proper smoothness assumption. A qualitatively very similar analysis of this is offered in Section 5.4.

2.1.6 Not Making Additional Assumptions

Ben-David et al. [1] provide a series of investigations starting from the conjecture that SSL is, in some sense, not possible without any additional distributional assumptions, such as those from Section 1.1. They hypothesize that, a semi-supervised learner cannot have essentially better sample complexity bounds than an SL (see Definitions 1 and 2). This setting is essentially different from the previous subsections, as there are no further restrictions on the model or the data generation process. In the following two subsections, we illustrate the precise idea of these conjectures. Additionally, we clarify why they do not hold generally and in which scenarios they are generally true. We start, however, with the main contributions from [1].

The generic hypothesis is that the worst-case sample complexity for any semi-supervised learner improves over a supervised learner at most by a constant that only depends on the hypothesis class. The first conjecture states this for the realizable case.

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

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

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

The second states the same for the agnostic case, i.e., we can replace D_h with any arbitrary distribution P.

Conjecture 2 ([1, Conjecture 5]). *For any hypothesis class H, there exists a constant $c(H)$ such that for any domain distribution D*

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

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

In other words, the paper conjectures that if we are given a fixed domain distribution, one can always find a labeling function (h in the realizable and $P(Y \mid X)$ in the agnostic case) for it such that the sample complexity gap between SL and SSL can only be a constant. The paper proves 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.

We note that the sample complexity comparison is, by construction, a worst case analysis. This means that in cases where the target hypothesis behaves benign, we could still get non-constant improvement. This is further explored in Section 5. On another note, we can also ask the question how good a constant improvement by itself can already be. We elaborate on this in the discussion section.

Conjectures 1 and 2 are both not true in full generality, which we will explain in the following subsections, but slightly modified statements may be shown.

In the *realizable case*, [21] shows that Conjecture 1 is true with a small alteration and when the hypothesis class has finite VC-dimension: if H is even finite, the supervised learner is allowed to be twice as inaccurate (note the 2ϵ in Inequality (3) below). If H is not finite but with finite VC-dimension, we get an additional term of $\log(\frac{1}{\epsilon})$ in Inequality (4). [22] takes this idea a step further and shows that there is a setting in which manifold regularization, which uses the manifold assumption, obeys the limits stated by the conjecture, even though in this case the domain distribution carries information about the labeling function. Specifically, [21] proves the following.

Theorem 1 ([21, 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 $h \in H$, if H is finite, then*

$$m(H, D_h, 2\epsilon, \delta) \leq O(\ln |H|) m^{\mathrm{SSL}}(H, D_h, \epsilon, \delta), \tag{3}$$

if H has finite VC-dimension, then

$$m(H, D_h, 2\epsilon, \delta) \leq O(\mathrm{VC}(H)) \log\left(\frac{1}{\epsilon}\right) m^{\mathrm{SSL}}(H, D_h, \epsilon, \delta). \tag{4}$$

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. Ref. [23] shows 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 [23, Proposition 4].

Regarding the *agnostic case*, Theorem 9 from [12] shows Conjecture 2 with some small modifications and assumptions. Like Theorem 1 it assumes finite VC-dimension together with further mild assumptions on the domain distribution D (while Conjecture 2 is formulated to hold for *all* distributions D). Another difference is that they consider an in-expectation and not a high-probability framework. The intuition for that result is straightforward: if we allow all labeling functions, i.e., consider the agnostic case, 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 semi-supervised learners.

In the case of a hypothesis class with infinite VC-dimension, however, both conjectures cease to hold, also for the slightly altered formulations. This is 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. We elaborate on this in the next subsection where we collect three different setups in which a semi-supervised learner can PAC-learn, while a supervised learner cannot.[3]

2.1.7 Not Restricting Possible Labeling Functions

We end with a related negative result from Ref. [23], which shows that if the domain $\mathcal{X}$ is finite and we allow all deterministic labeling functions on it, no semi-supervised learner can improve over a supervised learner that achieves 0 training error in the realizable PAC-learning framework, not even by a constant. The supervised learner is, however, to be allowed twice as inaccurate and twice as unsure, which is respectively captured by the 2ϵ and 2δ below.

Theorem 2 ([23, Theorem 8]). *Let $\mathcal{X}$ be a finite domain, and let $H_{\mathrm{all}} = \{0,1\}^{\mathcal{X}}$ be the set of all deterministic binary labeling functions on $\mathcal{X}$. Let A be any supervised learner that achieves 0 training error, P a distribution over $\mathcal{X}$ and $\epsilon, \delta \in (0,1)$. Then $m(A, H_{\mathrm{all}}, P, 2\epsilon, 2\delta) \leq m^{\mathrm{SSL}}(H_{\mathrm{all}}, P, \epsilon, \delta)$.*

3. Here, PAC-learnability means $m(H, \epsilon, \delta)$ is finite for all $\epsilon, \delta > 0$.

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 this setting one even loses this advantage. The idea of the result is similar to Theorem 9 from [12], discussed above: if there is no restriction on the labeling function it is difficult to learn the labels for the unobserved support.

In the next subsection, we see that positive results are still possible and present hypothesis classes on which semi-supervised learners can be effective. Following the previous result, it is not surprising, however, that those classes and the domain distributions they may operate on are carefully chosen.

2.2 On the Possibility of Semi-Supervised Learning

We consider three specific settings in which it can be shown that a semi-supervised learner can learn, while a SL cannot. We present the two works of [21] and [24], 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^* uniformly over the distributions of $\mathcal{D}^*$, while a semi-supervised learner that has access to the domain distribution can learn H^*. As a third, we present the work of [25] as we think it provides the most insightful example of how a shift from not learnable to learnable is possible when going from SL to SSL, even though in there we assume that the domain distribution restricts the possible labeling functions.

2.2.1 Proving the Realizable Case With a Discrete Set

Ref. [21] gives the first example that shows that Conjecture 1 does not hold in general. This is captured in the first theorem to follow.

Theorem 3 ([21, Theorem 2]). *There exists a hypothesis class H^* and a family of domain distributions $\mathcal{D}^*$ such that for every $D \in \mathcal{D}^*$,*

$$m^{\mathrm{SSL}}(H^*, D, \epsilon, \delta) \leq O\left(\frac{1}{\epsilon^2} + \frac{1}{\epsilon} \log\left(\frac{1}{\epsilon}\right)\right)$$

and, 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 semi-supervised learner 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.

Let us give some intuition for [21]'s example, which is also at the basis of the other results in this subsection. The setup is as follows. The domain $\mathcal{X}$ consists of all sequences

$$x = (x_1, x_2, \ldots, x_l)$$

of arbitrary finite length l and $x_i \in \{0,1\}$. The distributions $D \in \mathcal{D}^*$ on $\mathcal{X}$ are such that there is a sequence

$$D(x_{\sigma(1)} = 1) > D(x_{\sigma(2)} = 1) > \ldots,$$

where σ is a random permutation of the indices of x, and the distribution drops sufficiently quick in $\sigma(i)$.[4]

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 $\mathcal{D}^*$. After all, the VC-dimension could be finite over $\mathcal{D}^*$. We want to sketch how the semi-supervised learner 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 is good enough for predicting accurately. One then shows that the remaining hypotheses in H^* can be learned from finitely many samples.

The foregoing example, like those that follow, are essentially set up such that H and D have a certain link where knowledge about D can actually give knowledge about H. Note, however, that knowledge about D does not restrict the set of possible labeling functions from H, but it helps to identify which hypotheses can be safely ignored. Note also 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, there would be no position x_i which are with high probability 0 and we thus could not remove the corresponding hypotheses.

2.2.2 Proving the Agnostic Case Using Algebraic Varieties

Ref. [24] provides a different example of Theorem 3 for a continuous space $\mathcal{X}$, which may also be extended to the agnostic case, and thus refuting Conjecture 2 in full generality. Here the set of admissible distributions are given by specific manifolds. As such, they use the second, alternative, manifold assumption as given in Section 1.1.

Theorem 4 ([24, Theorem 5]). *There exists a hypothesis class H_{alg} and a set of distributions $\mathcal{D}_{\text{alg}}$ such that, for every $D \in \mathcal{D}_{\text{alg}}$,*

$$m^{\text{SSL}}(H_{\text{alg}}, D, \epsilon, \delta) < \frac{2}{\epsilon} \log \frac{2}{\delta}, \tag{5}$$

and the supervised sample complexity is infinite, i.e.,

$$\sup_{D \in \mathcal{D}_{\text{alg}}} m(H_{\text{alg}}, D, \epsilon, \delta) = \infty. \tag{6}$$

The hypothesis class H_{alg} consists of all hypotheses that have class label 1 on an algebraic set and 0 outside of that set. This algebraic set can essentially be considered a manifold of sorts. The hypotheses class is very rich and has infinite VC-dimension. If, however, we restrict the set of admissible domain distributions $\mathcal{D}_{\text{alg}}$ to be particular types 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 $V_1, \ldots, V_k$. Once we know that the distribution has support on $V_1, \ldots, V_k$, we only have to figure out which of those algebraic sets have label 1 and which have label 0. A semi-supervised learner can thus reduce the class H_{alg} by only considering the hypotheses that have class label 1 on combinations from $V_1, \ldots, V_k$. Since the set of all possible combinations is finite, a semi-supervised learner can learn them with a sample complexity bounded by Inequality (5).

The extension to the agnostic case might appear problematic at first, because the semi-supervised algorithm restricts the hypothesis set H_{alg}. To guarantee PAC-learnability, we need to know that the best predictor from 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 this, assume that the distribution is supported on one irreducible algebraic set V_0. Our semi-supervised learner can now choose to label it completely 1 or 0, where both options may lead to non-zero error. But labeling it completely as either 1 or 0 is already ideal, as using any algebraic set $V_1 \in H_{\text{alg}}$ will by construction be equal to V_0 (which leads to label everything as 1) or has an intersection of zero mass with V_0 (which leads to labeling almost everything as 0).

Interestingly, the findings above seems to contradict the results from Section 2.1.5. [19] shows that a supervised learner can also adapt to the underlying manifold. This discrepancy is explained by the fact that [19] restricts the target functions to be smooth, which presents the supervised learner with a sufficiently easy problem. The work in this section on the other hand confronts the supervised learner with an impossible, meaning not PAC-learnable, task.

2.2.3 Enforcing Learnability With the Manifold Assumption

Ref. [25] provides a third example in which a semi-supervised learner can effectively learn, while a supervised learner cannot. The motivation for this, however, was independent of [1] and meant as a general theoretical analysis of the manifold learning framework as introduced in [7]. Also, their results are in-expectation, while the previous papers give PAC bounds, i.e., they hold with high probability. The work relies on the manifold assumption, which limits the possible labeling functions, and thus is not a counterexample to 1. We believe, however, that it is the most intuitive setting to understand why a supervised learner cannot learn, while a semi-supervised learner can.

Though the paper presents the example in an in-expectation framework, we alter the setup slightly and present it in the PAC learning framework, which makes the comparison to the previous sections easier.

The example starts by assuming that the admissible domain distributions are given by the class of distributions $\mathcal{P}_c$ that have support on embeddings of a circle in the Euclidean plane (see Fig. 3). The hypothesis class H_c

4. Note that with $x_{\sigma(i)} = 1$ we mean the subset $V \subset \mathcal{X}$ with

$$V := \{x = (x_1, x_2, \ldots, x_l) \in \mathcal{X} | x_{\sigma(i)} = 1\}.$$

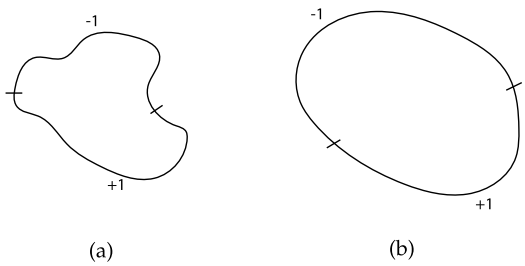


Fig. 3. The shapes shown in (a) and (b) are two different embeddings of a circle in the Euclidean plane. One half of the circle is labeled $+1$, while the other half is labeled as -1. Everything outside the circle is labeled $+1$.

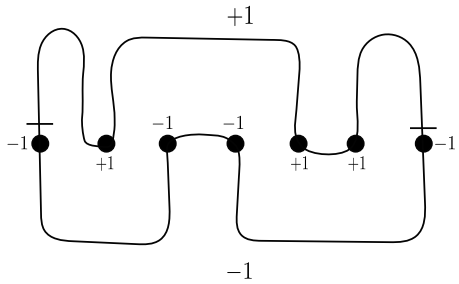


Fig. 4. A schematic proof why the hypothesis set H_c has an infinite VC-dimension. The embedded circle, its upper half assigning points to $+1$ and its lower to -1, can label the seven points correctly.

consists of all possible binary labelings of half circles, while everything outside the circle is labeled as 1.[5] The semi-supervised learner that knows the specific embedding of the circle only needs to find two thresholds on the given circle. This is a hypothesis class with a VC-dimension of 2, which implies that the semi-supervised learner can learn efficiently. In Fig. 4, we illustrate in a schematic way why H_c has an infinite VC-dimension and thus cannot be learned by any supervised learner.

3 Learning Without Assumptions

As we have seen in the previous section, it can be difficult to exploit unlabeled data not making additional assumptions. In fact, we saw that in various of these situations one can show that unlabeled data cannot help at all. As already mentioned in the introduction of Section 2, this impossibility sometimes stems from the fact that we only consider improvements of the estimate of the conditional probability $P(Y|X)$. This section looks at the complete risk $\mathbb{E}_{X,Y}[l(h(X), Y)]$, a quantity which is always influenced by the marginal distribution $P(X)$. Still, no additional assumptions about the distribution P are considered and the theoretical guarantees are weak accordingly.

Ref. [26] uses the unlabeled data to reweigh the labeled points and show improvements in terms of asymptotic efficiency. Interestingly, their result implies that strict improvements are only possible under model miss-specification. [27] employs the unlabeled data to determine the center of the version space. The best possible improvements in the learning rate as reported in that work are bounded by a factor of 2.

3.1 Reweighing Labeled Data by True Marginal

The work in [26] proposes a semi-supervised learner that has full knowledge of the marginal distribution $P(X)$ in a reweighing scheme, while $\mathcal{X}$ is assumed to be finite. [28] extends this to non-discrete features spaces. [26] considers models that directly estimate class probabilities $p(y|x, \theta)$, measuring performance by the negative log-likelihood

$$l(x, y|\theta) = -\ln p(y|x, \theta).$$

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

What is analyzed in the end is the asymptotic variance of the model estimation, in which two models are compared: the classical maximum log-likelihood estimate based on the labeled data only, i.e.,

$$\theta^{\mathrm{SL}} = \arg\min_{\theta\in\Theta} \sum_{(x,y)\in S_n} l(x, y|\theta), \tag{7}$$

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

$$\theta^{\mathrm{SSL}} = \arg\min_{\theta\in\Theta} \sum_{(x,y)\in S_n} \frac{P(x)}{\sum_{z\in X_n} I_{\{x=z\}}} l(x, y|\theta). \tag{8}$$

Note that the semi-supervised learner weighs each feature with the true, instead of the empirical, distribution.

Theorem 5 ([26, Theorem 1]). *Let*

$$\theta^* \in \arg\min_{\theta\in\Theta} \mathbb{E}[l(x, y|\theta)]$$

and define the following matrices

$$H(\theta^*) = \mathbb{E}_X\big[\mathbb{V}_{Y|X}[\nabla_\theta l(X, Y|\theta)|X]\big] \tag{9}$$

$$I(\theta^*) = \mathbb{E}_{X,Y}\big[\nabla_\theta l(X, Y|\theta)\nabla_\theta^T l(X, Y|\theta)\big] \tag{10}$$

$$J(\theta^*) = \mathbb{E}_{X,Y}\big[\nabla_\theta^T \nabla_\theta l(X, Y|\theta)\big], \tag{11}$$

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

$$\sqrt{n}(\theta^{\mathrm{SL}} - \theta^*) \to \mathcal{N}(0, J^{-1}(\theta^*)I(\theta^*)J^{-1}(\theta^*)) \tag{12}$$

$$\sqrt{n}(\theta^{\mathrm{SSL}} - \theta^*) \to \mathcal{N}(0, J^{-1}(\theta^*)H(\theta^*)J^{-1}(\theta^*)) \tag{13}$$

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

Asking now when θ^{SSL} dominates θ^{SL}, we get the surprising answer that this actually happens when the model is misspecified. It can certainly not happen, however, if the model is well-specified. In the latter case—along with some other regularity conditions, the MLE θ^{SL} is already asymptotically efficient. Moreover, we have that $H(\theta^*) = J(\theta^*) = I(\theta^*)$, which recovers the classical result that the MLE is asymptotically normal with a covariance that equals the inverse Fisher information matrix $I(\theta^*)$.

The paper examines, based on the logistic regression model, when the difference between $I(\theta^*)$ and $H(\theta^*)$ is particularly big and shows that this is the case the more

$P(Y|X)$ is bounded away from $1/2$, so in particular when the Bayes error is small. Such requirement on $P(Y|X)$ is very similar to the *Tsybakov-margin condition* [29], which is used in statistical learning to come to fast learning rates. In Sections 5.1 and 5.2, similar assumptions are presented based on which particular semi-supervised learners can converge exponentially fast to the Bayes error.

3.2 The Center of Version Space

Ref. [27] 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 really works in the realizable case. Although we do not get a new algorithm for the agnostic case, the paper presents novel bounds for supervised methods that make use of the unlabeled data.

3.2.1 Realizable Case

The idea for the realizable case is to consider the version space, i.e., 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 its hypotheses on this data. We are then going to take the hypothesis that has the lowest worst-case disagreement to all other hypothesis, amongst which must be the true hypothesis, as we assume realizability. Let us now make this more precise.

Given two hypotheses $f, g \in H$ we define the disagreement pseudo-metric $d(f, g)$ as

$$d(f, g) = P(f(X) \neq g(X)). \tag{14}$$

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

$$\hat{d}(f, g) = \frac{1}{m} \sum_{i=n}^{n+m} I_{\{f(x_i)=g(x_i)\}}. \tag{15}$$

The version space is defined as $H_0 = \{h \in H | \hat{R}(h) = 0\}$. If h_0 is 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, in turn, gets us to the following bound.

$$\begin{aligned} R(h) = d(h, h_0) &= \hat{d}(h, h_0) + (\hat{d} - d)(h, h_0) \\ &\leq \sup_{g \in H_0} \hat{d}(h, g) + \sup_{g, g' \in H_0} (\hat{d} - d)(g, g') \end{aligned}$$

As Inequality (16) 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 (16). More specifically, we choose the semi-supervised estimator to be the so-called *empirical center of the version space*:

$$h^{\mathrm{SSL}} = \arg \inf_{h \in H_0} \sup_{g \in H_0} \hat{d}(h, g). \tag{17}$$

With this we can of course only control the first term on the right-hand side of Inequality (16). In a standard way, we can bound the second term with concentration inequalities derived from a Rademacher complexity for the space

$$\mathcal{G} = \{x \mapsto I_{\{f(x)=g(x)\}} | f, g \in H_0\}.$$

Ultimately, this leads us to the result that with probability at least $1 - \delta$ [27, Theorem 3]

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

Note the two terms on the right-hand side of Inequality (18) go to 0 for increasing m and that, in this case, we also have that $\hat{d}(f, g) \to d(g, g)$. So, ignoring that we only have finitely many samples, we can compare the semi-supervised learner (17) to 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 additional information, we can always find a target hypothesis h^* such that

$$R(h^{\mathrm{SL}}) = \sup_{g \in H_0} d(h^{\mathrm{SL}}, g) = d(h^{\mathrm{SL}}, h^*).$$

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 come to the following bound:

$$R(h^{\mathrm{SSL}}) \leq \inf_{h \in H_0} \sup_{g \in H_0} d(h, g),$$

which holds at least for m going to infinity.

From a geometric viewpoint, $\sup_{g \in H_0} d(h^{\mathrm{SL}}, g)$ 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.

3.2.2 Bounds for the General Case

In the agnostic case, we do not assume that the target hypothesis is part of our hypothesis class. To still make use of the considered disagreement pseudo-metric to come to bounds, the author proposes the following general recipe.

The starting point is the observation that bounds for randomized classifiers are generally tighter compared to their deterministic counterparts [30], [31]. The idea is now to use such a randomized classifier f_{rand} as a kind of anchor. This anchor takes on a role similar to the target hypothesis in the realizable case. To get a bound for a classifier f, we can use the bound for the randomized classifier together with a slack term that includes $\hat{d}(f_{\mathrm{rand}}, f)$. 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. The paper additionally derives an explicit cross-validation bound, where the randomized classifier is given by a uniform distribution over the classifiers obtained in the multiple cross-validation rounds.

4 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. Here we investigate what can be achieved assuming, what we refer to as, *weak* assumptions. With weak assumptions we mean those that cannot essentially change the learning rate of $O(\frac{1}{\sqrt{n}})$, but rather give improvements by a constant which may depend on the hypothesis class. In Section 5, we investigate what we have to assume to actually escape the $\frac{1}{\sqrt{n}}$ regime.

We first cover the work of [32], as it provides a rather general framework that allows one to analyze the learning guarantees for various semi-supervised learners. This initial paper shows that semi-supervised learners that fall in this framework learn by a constant faster then supervised learners, where the constant depends on the hypothesis class and the semi-supervised learner considered. We then cover, in more detail, the idea of co-training. It can be studied within the framework of [32], but we present some additional details of interest not fully captured by this framework. In particular, we present the work in [33], that formulates the assumption of co-training in an information theoretical framework, which allows one to precisely quantify the bias-variance trade-off.

4.1 A General Framework for Weak Assumptions

Ref. [32] 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 [6], [34], multi-view assumptions [33], [35], and transductive graph-based methods [36] are just some examples. The idea is to introduce a function χ that measures the compatibility between a hypothesis h and the marginal distribution $P(X)$. For example, we can deem a hypothesis h compatible with a marginal distribution $P(X)$, if its decision boundary goes through low-density regions, encoding one assumption explained in Section 1.1.

Although χ should connect the marginal distribution $P(X)$ to the compatibility of a hypothesis h, it is much more useful to define χ for each point in the feature space individually. This way we can estimate χ based on a finite unlabeled sample, when we do not have access to the full distribution $P(X)$. Therefore χ is a mapping

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

The compatibility measure χ then gives rise to the function

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

which we refer to as the *unsupervised loss*. The aim is to optimize it in addition to the loss on the labeled sample.

Here, we focus on a single core theorem. The other results in the paper are similar in flavor and mostly differ in the realizability assumptions w.r.t. the unsupervised and the supervised error made and the bounding techniques employed. The paper presents bounds derived from uniform convergence as well as bounds based on covering numbers. The theorem presented considers the double agnostic case in which neither the labeled nor the unlabeled loss have to be zero.

Theorem 6 ([32, Theorem 10]). *Let*

$$h_t^* = \arg\min_{h \in H} [R(h) | R_{\text{unl}}(h) \leq t].$$

Then, given an unlabeled sample size of at least

$$\frac{64}{\epsilon_2^2}\left(2\max[VC(H), VC(\chi(H))]\ln\frac{1}{\epsilon_2} + \ln\frac{1}{\delta}\right),$$

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 $\hat{R}(h^{\text{SSL}})$ subject to $\hat{R}_{\text{unl}}(h^{\text{SSL}}) \leq t + \epsilon$, while $H(t) := \{h \in H | R_{\text{unl}}(h) \leq t\}$. Here $\hat{R}$ is the empirical risk measured with the sample S_n and $\hat{R}_{\text{unl}}$ is the empirical unlabeled risk measured on the sample U_m.

We note that the original paper uses (exponentiated) annealed entropy, see [5], instead of VC-dimension to measure complexity. To allow for an easier comparison to other results and avoid additional notation, we express the above theorem in terms of the standard VC-dimension. The difference of the latter measure is that it is distribution independent.

Let us briefly compare Theorem 6 to results from the previous section. In particular, let us consider 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 in Equation (21) by replacing $VC(H(t + 2\epsilon_2))$ with $VC(H)$. As we know that the complexity given by Equation (21) is tight up to some constants (see also [8], Chapter 6), 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 $\mathcal{O}(\frac{VC(H)}{VC(H(t+2\epsilon_2)})$. 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. It is exactly this type of example, as covered in Section 2.2, that refutes the conjecture.

Theorem 6 quantifies, to some degree, the fundamental bias-variance trade-off in SSL when we rely on additional 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 the compatibility function does not match the underlying problem however, we bias the procedure away from good solutions at the same time.

4.2 Assuming That the Feature Space can be Split

In multi-view learning, incidentally also referred to as 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, in principle, enough to learn. In the early work on co-training, [35] uses 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

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, and this disagreement can be measured with unlabeled data.

There are multiple theoretical results that pertain to this approach. It can, for example, be analyzed in the framework of the previous subsection. Alternatively, [37] and [38] analyze a Rademacher complexity term under the multi-view assumption, while [39] defines a kernel that directly includes the assumption as a regularization term, and thus find a RKHS where co-regularization automatically applies. Here, we detail the approach of [33] as it ties in best with the other results we present. In addition, [33]'s information theoretic framework allows us to also analyze the penalty one suffers if the assumption is not exactly true.

As above, we split the random variable X, which takes values in $\mathcal{X}$, into two: $X = (X^1, X^2)$. Now, the multi-view assumption from [33] can be formalized as follows: let $I(A; B|C)$ be the mutual information between random variables A and B, conditioned on the random variable C. We assume there exists an ϵ_{info} such that

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

and

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

In words: once we know one set of features, the other does not tell us much more about Y. Comparing this to co-training, we can see it as a relaxation: assuming that each view is already sufficient to fully learn, corresponds to an ϵ_{info} that equals 0. If, however, $\epsilon_{\text{info}} > 0$, we cannot learn perfectly from one view.

Subsequently, we assume that we have for each view X^1 and X^2 a corresponding hypothesis set H^1 and H^2. We carry out predictions with *pairs* of hypotheses

$$(f_1, f_2) \in H^1 \times H^2.$$

The paper uses the notion of compatibility functions, as generally defined through Equation (19). In particular, they define the compatibility function

$$\chi : H := H^1 \times H^2 \rightarrow [0, 1]$$

as

$$\chi(h^1, h^2, x) := d(f_1(x^1), f_2(x^2)),$$

where $d : \mathcal{Y} \times \mathcal{Y} \rightarrow [0, 1]$ is a specific pseudo-distance measure that fulfills a relaxed triangle inequality and $x = (x^1, x^2)$ is a sample. In essence, the distance d measures how much f_1 and f_2 agree on a sample x. For a given threshold $t \in \mathbb{R}$ we then find the best *pair* of hypotheses based on the empirical risk minimization problem

$$\min_{(h^1,h^2)\in H} \sum_{i=1}^{n} l(h^1(x_i^1), y_i) + l(h^2(x_i^2), y_i) \tag{24}$$

with additional constraint $\hat{R}_{\text{unl}}(h^1, h^2) \leq t$.

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 errors, corresponding with the loss l, when learning from $X^1 \times X^2$, X^1 and X^2, respectively. Also set

$$\epsilon_{\text{bayes}} = \max\{R(f_*^1) - \beta_*^1, R(f_*^2) - \beta_*^2\},$$

where f_*^i is the best predictor from H^i. Finally, recalling the definition of $R_{\text{unl}}(h)$ from Equation (20), define

$$\hat{H}(t) = \{(h^1, h^2) \in H | \hat{R}_{\text{unl}}(h^1, h^2) \leq t\}.$$

Theorem 7 ([33, Theorem 2]). [6] *Assuming the loss l is bounded by 1, there exists a $t \in \mathbb{R}$ (depending among others on ϵ_{info}, ϵ_{bayes}, and m), such that, if we have a labeled sample of size at least $m(\hat{H}(t), \epsilon, \delta)$, it holds with probability at least $1 - \delta$ that*

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

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

5 Learning Under Strong Assumptions

In the previous chapter, we analyzed assumptions that only could give improvements in terms of a multiplicative constant. These did not allow us to come to semi-supervised learners that improve beyond the general learning rate of $\frac{1}{\sqrt{n}}$. Here, we analyze assumptions, cf. Section 1.1, that enable us to escape this regime, even leading to exponentially fast convergence in some cases.

To illustrate how such improvements are possible, assume that one comes to a clustering based on all of the data provided and assume that this clustering is correct, i.e., each cluster corresponds mainly to one class. Under this assumption, we only need enough labeled data to identify which cluster belongs to which class, and this can be done exponentially fast. The work in the current 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? How can we go beyond classification and deal with regression?

5.1 Assuming the Model is Identifiable

One of the classic analyses in semi-supervised learning deals with identifiable mixture models and deals with a particular notion of sample complexity [40]. As it turns out, the setting is quite restrictive but can, as such, give exponentially fast convergence to the Bayes risk. The outcome is very strong, considering that the results covered in the previous sections were essentially unable to improve upon the standard convergence rate of $\frac{1}{\sqrt{n}}$. Consider for instance Inequality (21) after solving for ϵ.

The first key assumption to actually obtain these results lies in the data generation process. First, the label is drawn with $P(y = 1) = \eta$ and $P(y = 0) = \bar{\eta}$. Then a feature vector is drawn according to a density $f_y(x)$. Unlabeled data is

6. The theorem actually needs some additional regularity conditions. These are not made explicit to aid in focusing on the main point.

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., we can infer the mixture model uniquely given enough unlabeled data. After identifying the mixture, we merely have to figure out how to label each part of the two mixture components. Deciding between the remaining alternatives can be done by a simple likelihood ratio test, which converges exponentially fast to the Bayes risk in the number of labeled samples n:

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

For the analysis it is necessary to assume that one has an infinite amount of unlabeled data. The work is continued in [41], where the authors consider cases where we already have knowledge about the densities f_y. [42] considers a similar framework for the case where the marginal distribution $P(x)$ is unknown, and instead assume that $P(x)$ can be well estimated with a mixture of two spherical Gaussian distributions.

The above work ties in with the impossibility result from Section 2.1.1. Here, however, the data generating process is reversed: the feature x depends on y and thus violates the data generation of Fig. 1 from 2.1.1, which led to an impossibility result.

5.2 Assuming Classes are Clustered and Separated

Reference [43] presents explicit bounds on the generalization error using an alternative formulation of the cluster assumption. The approach closely resembles the work described in the previous subsection and, similarly, enables exponentially fast convergence under semi-supervision.

The work's initial, elementary setup is that we are given a collection of pairwise disjoint clusters $C_1, C_2, \ldots$ for which we assume that the optimal labeling function

$$x \mapsto \operatorname{sign}\left(P(Y = 1 | X = x) - \frac{1}{2} \right)$$

is constant on each cluster C_i. So the clusters have a label-purity of some degree, which we can express as follows:

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

The cluster C_i is called pure if and only if $\delta_i = 1$.

Assuming now that we know the clusters, we let $h_n^{\mathrm{SSL}}(x)$ be the majority voting classifier per cluster. More formally, given a labeled sample S_n let $X_i^+ := \{(x, y) \in S_n | x \in C_i, y = 1\}$ and similarly $X_i^- := \{(x, y) \in S_n | x \in C_i, y = -1\}$. Then given a new data point $x \in C_i$ we set

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

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, avoiding the slow rates that we otherwise may obtain as explained in the penultimate paragraph from Section 2.1.6. Thus the excess risk of interest is restricted to the set $C := \bigcup C_i$ and so we consider the risk

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

where h^* is the Bayes classifier. The following theorem expresses the possible gain with respect to the expected cluster excess risk.

Theorem 8 ([43, 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*

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

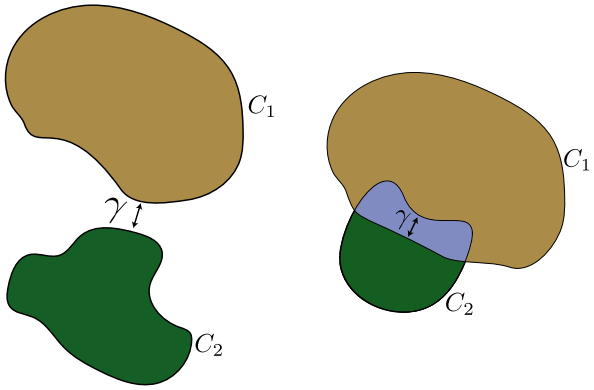
That is, knowing the clusters, we recover the exponential convergence in the labeled sample size as in Section 5.1.

The biggest effort of the paper goes into the definition of clusters and the finite sample size estimation of such. The derivations are rather extensive and, as in most of the review, we limit ourselves here to a description of the underlying intuition. To start with, one assumes that the marginal distribution $P(X)$ allows for a density function $p(x)$. One can then define the density level sets of $\mathcal{X}$ w.r.t. a parameter $\lambda > 0$ as $\Gamma(\lambda) := \{x \in \mathcal{X} \,|\, p(x) \geq \lambda\}$. 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. Therefore, to ensure that we still can learn exponentially fast, the number of unlabeled samples has to grow exponentially with the number of labeled samples.

Finally note that the previous analysis is not a PAC-analysis: the result in Inequality (29) is not over a worst case distribution. Performing such worst case analysis, we may for a given n chose a distribution with $\delta_i = \sqrt{\frac{1}{n}}$. Plugging this δ_i into Inequality (29), we observe that the exponential rate actually turns into a slow rate (cf. Fig. 1 from [44] for a similar observation). One way to avoid this problem is to assume that the posterior distribution $P(Y \,|\, X)$ is bounded away from $\frac{1}{2}$, which directly implies that we cannot chose δ_i as above. Consequently, one may wonder if the PAC-framework isn't overly pessimistic, which is a topic we return to briefly in the discussion.

5.3 Classes Clustered but not Necessarily Separated

Reference [45] propose yet another formalization of the cluster assumption. More specifically, it is one that allows to distinguish cases where SSL does help and where not. This is achieved 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 constitute of different clusters that are at times easy to distinguish and in other cases not. The marginal densities $p(x)$ from $\mathcal{P}$ are given by



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

(b) Light blue is the cluster overlap with margin γ. The three colors constitute three different decision sets.

Fig. 5. The idea of (a) a positive and (b) a negative γ-margin.

mixtures of K densities p_k. That is, $p(x) = \sum_{i=1}^{K} a_k p_k(x)$ 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 particular regularity conditions. We refer to these sets C_k as clusters and each of these is assumed to have its own smooth label distribution function $p_k(y|x)$. So with probability a_k we draw from $p_k(x)$ and then label x according to $p_k(y|x)$. We further only consider distributions that lead to clusters with margin, with our without overlap, of at least γ (see also Fig. 5), and denote the resulting class of distributions by $\mathcal{P}(\gamma)$.

The clusters are not the main interest, but rather what the authors call the *decision sets*. To define a decision set, we take C_k^c to be the complement of C_k and, in addition, define $C_k^{\neg c} := C_k$. Now, a set $D \subset \mathcal{X}$ is called a decision set if it can be written as $D = \bigcap_{k \in K} C_k^{i_k}$ with $i_k \in \{c, \neg c\}$ for all $k \in K$. See Fig. 5b for an example. On the decision sets $p(x, y)$ is smooth as long as each $p_k(y|x)$ is smooth, while $p(x, y)$ is not necessarily smooth on each cluster, as it might exhibit jumps at the borders. Consequently, knowing the decision sets, one can use a semi-supervised learner that exploits the smoothness assumption.

The main theorem answers the question whether one can learn the decision sets from finitely many unlabeled points, which is done with the help of a marginal density estimator whose spacing is proportional to a parameter κ_0.

Theorem 9 ([45, 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 by $\mathcal{E}_{\max}$ and that there is a learner h_n^D that has knowledge of all decision sets D and, additionally, fulfills the 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 $|\gamma| > 6\sqrt{d}\kappa_0 (\frac{(\ln m)^2}{m})^{\frac{1}{d}}$, then there exists an $h_{n,m}^{\text{SSL}}$ such that

$$\sup_{P \in \mathcal{P}(\gamma)} \mathbb{E}_P[\mathcal{E}(h_{n,m}^{SS})] \leq \epsilon_2(n) + \mathcal{E}_{\max}\left(\frac{1}{m} + 2\kappa_0 C\sqrt{d}(n+1)\left(\frac{(\ln m)^2}{m}\right)^{\frac{1}{d}}\right), \tag{31}$$

where $C \geq 1$ is a constant that depends on smoothness properties of the boundary of the decision sets.

We immediately 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 exponentially more unlabeled than labeled data, which corresponds exactly to the finding in the previous subsection.

All in all, the intuition behind the theorem is fairly straightforward. The bigger γ, the less unlabeled samples we need to estimate the decision sets D. Moreover, once we know those sets, we can perform as well as h_n^D. Now, to analyze if a semi-supervised learner that first learns the decision sets empirically has an advantage over all supervised learners, we first find minimax lower bounds for all fully supervised learners. We can then give upper bounds for a specific semi-supervised learner and the conclusions follow easily: 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. As a consequence, γ should not be too big, because then the supervised learner can also distinguish the decision sets. Of course, the unlabeled data should not be too small, for then the semi-supervised learner cannot distinguish the decision sets either.

To showcase 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 cases where 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 semi-supervised learner is upper bounded by $n^{-\frac{2\alpha}{2\alpha+d}}$.

5.4 Smooth Regression Along a Manifold

As we elaborate on in the discussion section, an issue in SSL is that most methods are based on assumptions on the full distribution. The core problem is that we usually cannot verify whether such assumptions hold 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 [46] is one of the few papers that touches on this topic and introduces 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 gives a formalization of the manifold assumption and uses regression, while most others deal with classification, we decided for a fairly detailed presentation.

The authors use a version of the manifold assumption, so we enforce our estimated regression function $h^{\mathrm{SSL}}(x)$ to behave smoothly in high density regions. The density of the marginal distribution $P(X)$ is measured with a smoothed density function $p_\sigma(x)$

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

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 on $\mathbb{R}^d$, i.e., the so-called α, σ-exponential metric, that depends on an $\alpha \geq 0$ 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. This behavior gets of course more 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 a semi-supervised estimator as follows:

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

The estimator is a nearest-neighbor regressor, where neighbors are weighted according to their distance in terms of the previously defined exponential metric. The manifold assumption is employed by restricting the analysis to a class of distributions, $P(\alpha, \sigma, L)$, which only contains distributions such the regression task is L-Lipschitz w.r.t. the α, σ-exponential metric.

The following theorems gives bounds on the squared risk of h^{SSL} under the assumption that $\sup_{y \in \mathcal{Y}} |y| = M < \infty$.

Theorem 10 ([46, Theorem 4.1]). *Assume we have an unlabeled sample size m large enough to ensure that that for all $P \in P(\alpha, \sigma, L)$, $P(||\hat{p}_\sigma - p_\sigma|| \geq \epsilon_m) \leq 1/m$. Then*

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

Here, $\mathcal{N}_{P,\alpha,\sigma}(\epsilon)$ is the *covering number* of P in the exponential metric, i.e., the minimum number of closed balls in $\mathcal{X}$ of size ϵ (w.r.t. to the exponential metric) necessary to cover the support of $P(X)$ (see also [8], Chapter 27). In the Euclidean case, when $\alpha = 0$, we can bound $\mathcal{N}_{P,\alpha,\sigma}(\epsilon) \leq (\frac{C}{\epsilon})^d$ with the help of 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 previous theorem may be difficult to grasp in full at a first read and the paper offers, under some further regularity conditions, a simplified corollary in addition.

Corollary 1 ([46, Corollary 4.2]). *Assume that $\mathcal{N}_{P,\alpha,\sigma}(\delta) \leq (\frac{C}{\delta})^\xi$ for a certain range of δ. Furthermore, assume that m is large enough and that τ decreases at an appropriate rate, depending on n, α and ξ.[7] Then for all $P \in \mathcal{P}(\alpha, \sigma, L)$ the following inequality holds asymptotically and up to constants*

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

Following this, the paper analyzes the additional penalty one occurs by trying to find the best α. We start 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 now that we have, in addition to the training sample S_n, also a validation set $V = \{(v_1, z_1), \ldots, (v_n, z_n)\}$, which, for convenience, is also of size n. Let h_θ^{SSL} be the semi-supervised hypothesis trained on S_n with parameters θ. We then choose the final hypothesis h^{SSL} by optimizing for θ on the validation set:

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

Theorem 11 ([46, Theorem 6.1]). *Let $\mathcal{E}(h) := R(h) - R(h^*)$ be the excess risk, where h^* is the true regression function. There exist constants[8] $0 < a < 1$ and $0 < t < \frac{15}{38(M^2+\sigma^2)}$ such that*

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

This result is particularly interesting since we can always compare implicitly 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 size $O(\frac{\ln(n)}{n})$. This of course allows us to flexibly choose between the semi-supervised and the supervised method.

In a final contribution, the authors identify a case where the semi-supervised learning rate can be strictly better than the supervised learning rate. The setting considered is much like the one we have seen in Section 2.2. In particular, they construct a set of distributions $\mathcal{P}_n$, which depends on the number of labeled samples, such that 1) the estimator $h^{\mathrm{SSL}}(x)_{\tau,\alpha,\sigma}$, as defined in Equation (34), fulfills

$$\sup_{P \in \mathcal{P}_n} \mathbb{E}_{S_n}[R(h^{\cdot 2SSL})] \leq \Big(\frac{C}{n}\Big)^{\frac{2}{2+\xi}},$$

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

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

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 $d - 3$ [46, Lemma 1]. Worth noting is that the construction of $\mathcal{P}_n$ works by concentrating

7. This rate is specified in the actual paper.

8. It should be noted that these are not universal. They depend to some degree on the problem at hand.

the distributions more for larger 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 5.3, 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 2.2.3: if we would restrict the domain distributions such that only smooth circle embeddings would be allowed, a supervised learner could also learn efficiently as then a finite number of labeled samples would be sufficient to learn the domain distribution uniformly.

6 Learning in the Transductive Setting

SSL methods use unlabeled data to try and find better inductive classification rules, i.e., rules that apply to the whole input domain $\mathcal{X}$. Some works, however, consider schemes where one only cares about the labels of the unlabeled data specifically at hand. Such methods are often referred to as transductive and have been argued to be an essential step forward compared to inductive methods, in particular by Vapnik (see, for instance, [5, Chapter 8] and [4, Chapter 25]. While we review the most important theoretical results, a more detailed overview can be found in Chapter 2 of [47]. In Section 6.1, we present learning bounds that apply specifically to this transductive setting, though they often arise as direct extensions to the *supervised* inductive case. In Section 6.2, we present two papers that touch on the topic of so-called safe semi-supervised learners[9], where one constructs semi-supervised learners that are never worse than their supervised counterparts.

One essential difference, based on which two distinct transductive settings can be identified, is the way the sampling of the labeled and unlabeled data comes about.

Setting 1.

1) We start with a fixed set of points $X_{n+m} = \{x_1, \ldots, x_{n+m}\}$.
2) We reveal the labels Y_n of a subset $X_n \subset X_{n+m}$, which is uniformly selected at random. For notational convenience and without loss of generality, we usually assume 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 aim to find a classifier h with good performance as given by $R_m(h) := \sum_{i=n+1}^{n+m} l(h(x_i), y_i)$.

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 an additional 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 as specified by $\mathbb{E}_{S_n, X_m}[\frac{1}{m}\sum_{i=n+1}^{n+m} l(h(x_i), y_i)]$.

The work we present here deals with Setting 1. This is primarily out of convenience, but we note that one can always transform bounds from Setting 1 to bounds in Setting 2 [5, Theorem 8.1].

9. Incidentally, this is a topic that is not covered in [47].

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

6.1 Transductive Learning Bounds

The study of transductive inference goes back at least to the original work by Vapnik [48]. In this subsection, our primary source is [5] and we mainly consider the result found as Equation (8.15) in Theorem 8.2 of that work.

Assume that we are given $n + m$ samples and we take at random n samples on which to train. We then want to estimate the error on the remaining m samples. Vapnik shows that a hypergeometric distribution describes the probability that the observed error on the train and test set is larger than ϵ. 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[10] 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 ([5, Eq. (8.15)]). *For all $h \in \{-1, 1\}^{n+m}$, the following inequality holds with a probability of $1 - \delta$:*

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

A core problem with this inequality is that the term ϵ^* is an implicit function of n, m, δ and h and, as such, it is unclear what the learning rates are that we can actually achieve. The paper addressing this issue is covered next.

6.1.1 Bounds as a Direct Extension of Inductive Bounds

Ref. [49] finds explicit transductive bounds in a PAC-Bayes framework. We present a bound from the paper that is essentially a direct extension of a supervised inductive bound from [50]. Their result considers a Gibbs classifier, which we first introduce.

Let q be any distribution over the H. The Gibbs classifier G_q classifies a new instance $x \in \mathcal{X}$ with an $h \in H$ drawn according to q. The risk of G_q over the set S_n is then

$$R_n(G_q) = \mathbb{E}_{h \sim q}\left[\frac{1}{n}\sum_{i=1}^{n} l(h(x_i), y_i)\right].$$

Theorem 13 ([49, 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*

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

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

This theorem is indeed a direct extension of the inductive supervised case as found under Equation (6) in [50]. The only difference is that the term $\frac{m+n}{m}$ is missing. Although [51] shows 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, however, 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 [52].

[49] exploits the previous observation 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 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, one should 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 4 and 5. Using the clustering approach sketched above reduces the penalty in Inequality (40), which is coming from $\mathrm{KL}(q||p)$. In other words: we reduce the variance of the classifier. Clearly, on the other hand, using a clustering approach biases our solution and we get degraded performance compared to a supervised solution if the clusterings have a high impurity, i.e., clusters do not have clear majority classes.

6.1.2 Bounds Based on Stability

In [53], transductive bounds are explored under the assumption of stability, i.e., the notion that the output of a classifier does not change much if we perturb the input a bit. The transductive bounds presented are an extension of the inductive bounds that use the notion of *uniform stability* (see [54]) and *weak stability* (see [55], [56]). We cover the simpler transductive bound based on uniform stability and explain the difference to weak stability.

Assume that $h^{\mathrm{trans}} \in H$ is a transductive learner. That is, a hypothesis that we (deterministically) choose based on a labeled set S_n and an unlabeled set X_m. Furthermore, define $S_n^{ij} := (S_n \setminus \{(x_i, y_i)\}) \cup \{(x_j, y_j)\}$. 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. Similarly, define $X_m^{ij} := (X_m \setminus \{x_j\}) \cup \{x_i\}$. 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

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

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. With $\gamma > 0$, we define

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

Consequently, we can 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 ([53, Thereom 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*

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

Note that β is depended on n and m and we expect that the bigger our training set is, the less our algorithm changes if we exchange two samples from the train and test set. The transductive bounds based on Rademacher complexities, reviewed in the next subsection, can achieve convergence rates of $\frac{1}{\sqrt{\min(m,n)}}$. To obtain the same rate with Inequality (43), we need that β behaves as $O(\sqrt{(\frac{1}{n}+\frac{1}{m})\frac{1}{\min(n,m)}})$. This stability rate can indeed be achieved for regularized RKHS methods as demonstrated in [57] for laplacian normalization for graph-based SSL.

6.1.3 Transductive Rademacher Complexities

Rademacher complexities are a well studied and established tool for risk bounds in the inductive case [58]. [59] 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 H data-dependent. The definition of the transductive Rademacher complexity of a hypothesis class H closely follows the inductive case and is 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 (Theorem 6) for all $h \in H$, we have that with probability of at least $1-\delta$

$$R_m(h) \leq R_n^\gamma(h) + \frac{\mathrm{tRad}(H)}{\gamma} + \left(\frac{2}{\sqrt{\min(m,n)}}\right)\left(\sqrt{\frac{32\ln(4e)}{3}} + \sqrt{\frac{8}{3}\ln\left(\frac{1}{\delta}\right)}\right).$$

This bound can be used to directly estimate the transductive risk for transductive algorithms.

At a first glance, the inequality may seem somewhat surprising considering that the labeled and unlabeled data play an equivalent role in terms of convergence. While slow convergence for $n \ll m$ may be expected, one has to realize that, in case $m \ll n$, the transductive risk has very high

variance and therefore large intervals for high-confidence estimations are obtained.

[60] makes different use of Rademacher complexities in their derivation of 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 a certain way compatible with the clusters obtained. 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 $\tilde{O}(\frac{\sqrt{K}}{\sqrt{n}} + \frac{K^{3/2}}{\sqrt{m}})$ (see [60], Corollary 4). Not surprisingly, the learning rates are essentially the same as in the binary transductive cases. We note, however, that the analysis was done within Setting 2.

6.1.4 Bounds Based on Learning a Kernel

As a direct extension of the inductive case (see, for example, [61]), [62] proposes 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 example kernel classes are designed as follows. Given an initial set of kernels $\{K_1, \ldots, K_k\}$, define

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

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

Restricting the trace of the kernels allows us to bound later the complexity of the following defined hypothesis set.

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

We now come to the paper's claim, which is a bound on the transductive risk when using the above hypothesis set. The original formulation of the theorem is rather long and contains some additional definitions and clarification as part of it. In an attempt to make the presentation easier to access, we formulate the core result as a theorem, which should convey its basic structure and idea. Only afterwards, we will provide the missing details of the theorem.

Theorem 15 ([62, Theorem 24]). *For every $\gamma > 0$, with probability at least $1 - \delta$ over every training and test set of size n (so $m = n$), uniformly chosen from (X, Y), we have for every $h \in H_{\mathcal{K}}$:*

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

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

This last measure of complexity, $\mathcal{K}$, is 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. For the previously defined kernel classes $\mathcal{K}_c$ and $\mathcal{K}_c^+$, this complexity measure can, in turn, be bounded by

$$\mathcal{K}_c = c\mathbb{E}\left[\max_{K \in \mathcal{K}} \sigma^t \frac{K}{\mathrm{trace} K} \sigma \right] \leq cn,$$

and

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

In this last expression, λ_j is the largest eigenvalue of K_j.

Since m is taken to equal n, we find that the above bound gives the same learning rate $O(\frac{1}{\sqrt{m+n}})$ as we also found in Sections 6.1.2 and 6.1.3. We would, however, not expect that the rate of $O(\frac{1}{\sqrt{m+n}})$ also holds for different choices of m and n. We would rather expect to find $O(\frac{1}{\sqrt{\min(m,n)}})$, for the same reason as in the previous subsection (high variance of test risk for small m).

On another note, we point out that the effect the unlabeled data has on this procedure depends on the initial kernel guesses $\{K_1, \ldots, K_k\}$, but the choice of the kernels is actually independent of the unlabeled data. The unlabeled data may, however, inform the choice for the kernels as proposed in [63] and [64]. In essence, these works construct kernels that encode a manifold assumption by constraining the kernels to be smooth with respect to a given graph-structure of the unlabeled data.

6.2 Safe Transductive Learning

In the semi-supervised learning community, it is well known that using a semi-supervised procedure comes with a risk of performance degradation [2]. This problem leads some authors to ask the question whether it is possible to perform semi-supervised learning in a safe way: can one guarantee that the semi-supervised learner is not worse than its supervised counterpart. We note that, for risk bounds, a smaller bound still does not guarantee improvement, even if the underlying assumptions are correct.

We look specifically at the approaches from [65] and [17]. The results from both works are based on a minimax formulation and show that, in certain settings, one can indeed get to guarantee performance improvements by using SSL. The analysis is done in transductive Setting 1, which means that we have a training set S_n and a test set X_m.

6.2.1 A Minimax Approach for SVMs

The baseline for the model proposed in [65] is the $S3VM$ [66], which takes the unlabeled data into account by finding a large-margin solution. The proposed model $S4VM$ finds a few diverse proposal large-margin solutions, and then picks amongst these by means of a minimax framework to hedge against possible worst case scenarios. The idea is that, given that we found a set of a few potential solutions $H_p = \{h_1, \ldots, h_T\}$, we compare those solutions to h^{SVM} and then

choose the one with the biggest gain over h^{SVM} within a minimax framework.

Assume for now that we know the true labels $Y_m = (y_n, \ldots, y_{n+m})$ 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:

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

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

Defining our objective to be the difference, i.e., $J(h, y, h^{SVM}) = \text{gain}(h, Y_m, h^{SVM}) - \text{loss}(h, Y_m, h^{SVM})$, we can define a semi-supervised model h^{SSL} as the maximizer of this difference. The problem is, of course, that we actually do not know the true labels. Therefore, let us assume a worst-case scenario, which leads us to the following max-min formulation:

$$h^{\text{SSL}} = \arg\max_{h\in H_p} \min_{Y\in Y_p} J(h, Y, h^{SVM}). \quad (46)$$

Here $Y_p = \{(h(u_1), \ldots, h(u_m))|h \in H_p\}$ is the set of all possible labelings that we can achieve with H_p. To guarantee that our semi-supervised learner is not worse than the supervised learner it is important to assume that the true labels Y_m are part of the set Y_p, because only then we can guarantee what follows.

Theorem 16 ([65, 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.*

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 through Y_p, i.e, the set of all labelings that we consider possible. While this result still relies on some assumptions, [17] gives a case of guaranteed assumption-free improvements. This, however, comes at the cost of measuring improvement in terms of likelihood, and not accuracy, as described in what follows.

6.2.2 A Minimax Approach for Generative Models

The technique taken from [17], also in the line of safe SSL research, is, to our knowledge, the only approach to semi-supervised learning that considers a completely assumption-free setting. This comes at a cost, of course, which we will expand on later.

The starting point is a family of probability density functions $p(x, y|\theta)$ on $\mathcal{X} \times \mathcal{Y}$, where $\theta \in \Theta$ is a parametrization. We then fix θ^{SL} to be the supervised maximum likelihood estimator for the model $p(x, y|\theta)$, i.e.,

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

Let us assume for now that we know the true conditional probabilities $p = (p_1, \ldots, p_{m+n}) \in [0, 1]^{m+n}$ with $p_i = p(1|x_i)$ for $x_i \in S_n \cup X_m$. Indeed knowing this, we would rather optimize the expected log-likelihood of the model $p(x, y|\theta)$ evaluated on the complete dataset $X_{n+m} = \{x_1, \ldots, x_{n+m}\}$. This likelihood is given by

$$L(\theta|X_{n+m}, p) = \mathbb{E}_{Y\sim p}\left[\sum_{x\in X_{n+m}} \ln p(x, Y|\theta) \right]. \quad (47)$$

To be better than the supervised model θ^{SL} on the complete (transductive) likelihood in Equation (47), we would like to maximize the likelihood gain over it. In other words, we want to find the θ that maximizes the difference

$$C(\theta, \theta^{\text{SL}}|X_{n+m}, p) = L(\theta|X_{n+m}, p) - L(\theta^{\text{SL}}|X_{n+m}, p). \quad (48)$$

Clearly, we cannot maximize (48) directly, as we do not know the true probabilities p. Take, however, $p(y_i|x_i) = 1$ for all labeled points $(x_i, y_i) \in S_n$ and set $p_n = (p(1|x_1), \ldots, p(1|x_n)) \in \{0, 1\}^n$. For the unlabeled points X_m, we assume worst case posteriors denoted by the m-vector p_m, and consider the following max-min formulation:

$$\theta^{\text{SSL}} = \arg\max_{\theta\in\Theta} \min_{p_m\in[0,1]^m} C(\theta, \theta^{\text{SL}}|X_{n+m}, (p_n, p_m)). \quad (49)$$

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

Theorem 17 ([17, Lemma 1]). *Let θ^{SSL} be a solution found in Equation (49), then $L(\theta^{\text{SL}}|X_{n+m}, Y_{n+m}) \leq L(\theta^{\text{SSL}}|X_{n+m}, Y_{n+m})$.*

Subsequently, [17] shows that for some specific choices for the model $p(x, y|\theta)$, the previous inequality is strict almost surely, i.e., with probability 1 and we are guaranteed that the transductive likelihood of our semi-supervised model is larger than that of the supervised model. [67] proofs similar results for the least squares classifier using projection estimators.

An important difference between this work and the one from the previous subsection is that here one employs a generative model $p(x, y)$, while the SVM used in [65] is a discriminative model that inherently optimizes the class probability $p(y|x)$. The work in [18] (see also Section 2.1.4) shows that, to some degree, it is actually necessary to use a generative model as the semi-supervised estimator of Equation (49) coincides 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, which is a quantity that can be measured in part from unlabeled data.

One can imagine to use the minimax concept of this section also in the framework explained in Section 4.1, which uses any type of unlabeled loss. Note that the generative model of this section can always be decomposed into a class probability and a marginal distribution, which strongly

resembles the decomposition into labeled and unlabeled loss of Section 4.1.

7 Discussion and Conclusion

We comprehensively surveyed the theory that informs us about the potential of semi-supervised learning for improvements and the possible lack of it. Wrapping up, we point out some issues that, we believe, get to the core of the matter.

7.1 On the Limits of Assumption Free SSL

In Section 2, we reviewed work that analyzes the limits of semi-supervised learning when no particular assumptions about the distribution are made that a semi-supervised learner can exploit. The most general formulation is captured in Conjectures 1 and 2. They essentially state that a semi-supervised learner can beat all supervised learners by at most a constant. We then cover work that shows that the conjectures do actually not hold generally for hypothesis spaces of infinite VC-dimension. They do hold for finite VC-dimension spaces, but only under further relaxations.

7.2 How Good can Constant Improvement Be?

The question studied Section 2.1.6 is whether a semi-supervised learner can offer more than a constant improvement in terms of sample complexity. It seems equally fair, however, to ask how good a constant improvement can be. It is at least something that, certainly from a practical point of view, could still be very beneficial. The answer can be obtained through a thought experiment.

Assume that we have two classes given by two concentric d-dimensional spheres. Even if we assume that we have enough unlabeled data for a manifold regularization scheme to identify the spheres, we know that manifold regularization can only achieve constant improvement [22]. The intuitive explanation is the same as for Theorem 9 from [12] as explained in Section 2.1.6: if we allow arbitrary noise levels on the spheres we can only learn with same slow rate of any supervised learner, so we may only get constant improvements.

This constant, however, can be arbitrarily large. 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. [22] shows 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 $\mathrm{VC}(H)$ increases with the dimension d. This shows that the constant improvement can be arbitrarily large. While this example uses the manifold assumption, [23] gives an example, cf. Section 2.1.6, with a semi-supervised learner that has full knowledge of the domain distribution.

All in all, this shows that constant improvement can be arbitrarily large under the right assumptions, e.g., the manifold assumption, or full knowledge of the marginal distribution. An open problem that we identify is whether one can also have arbitrarily high constants with limited unlabeled and data without assumptions.

7.3 The Amount of Unlabeled Data We Need

In Section 2.2, we treated 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. 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 [12, Theorem 1]. The way the examples of Section 2.2 work is that for each fixed P a finite, bur arbitrarily large, amount of unlabeled data is sufficient. As a consequence, if we want to learn over all possible distributions, we need an arbitrarily large, i.e., infinite, amount of unlabeled data.

The semi-supervised improvements which we presented in Sections 3, 4 and 5, do not necessarily need an infinite amount of unlabeled data, although this is sometimes assumed for convenience. The difference is that, in those settings, supervised learners are also able to PAC-learn, but a semi-supervised learner is able to do this with fewer labeled samples. In Sections 5.2 and 5.3, we saw two instantiations of the cluster assumption where, to exploit them, the amount of unlabeled data needs to increase exponentially with the amount of labeled data. This is because the error in finding the clusters decreases only polynomially in the number of unlabeled points as shown in Inequality 31.

Having a finite amount of unlabeled data turns out to be surprisingly restrictive. In the light of the previous results, we believe that any limit on unlabeled data prevents us from proving results that hold uniformly over all data distributions. Identifying settings where a such limited amount leads to large (constant) improvements remains an open and challenging problem. We note that a positive result such as Theorem 3 is impossible with a fixed finite amount of unlabeled data [12, Theorem 7]. One may thus wonder what the strongest possible results are in the setting with a fixed amount of unlabeled data.

7.4 Assumptions in Semi-Supervised Learning

In Sections 4 and 5, we investigated what a semi-supervised learner can achieve once assumptions such as those from Section 1.1 are made. Any such 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 problem with this is that we do not know if such assumptions do hold or not. Clearly, one may be able to test the validity of certain assumption, but we conjecture that testing for an assumption consumes as many labeled points as learning directly a good classification rule with a supervised learner. In other words, the test would defy its purpose.

To make this claim a bit more precise, let us 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 if and only if $P \in \mathcal{P}^A$. For example, $\mathcal{P}^A$ could only contain distributions such that the marginal distributions $P(X)$ have always support on clusters and each cluster has a unique label. The important thing to note is

that assumption A is a property on P, so we need labeled samples to test its validity. 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 5.4. [46] shows that one can get essentially faster rates if the assumption is true, but we pay a penalty of $O(\ln(n)/n)$ if it is not true. [68] investigates how one can test for a property in an active way, so when we can choose which samples we want to label. Analyzing the assumptions made of different SSL methods this way could shed more light on their applicability. The implications of this testing procedure for semi-supervised learning are, at this point, a further open research question. Of course, one could insist that it is just not necessary to test whether an assumption is true or not. Following Vapnik's motto, we may want to avoid any intermediate form of testing to decide if an assumption is true or not, when, ultimately, we are merely interested in whether the semi-supervised learner performs better or not. Investigating whether semi-supervised learning is only possible with prior knowledge is thus an further interesting open problem.

7.5 Weak Versus Strong Assumptions

Distinguishing between weak and strong assumptions can be motivated through their (in)ability to improve the learning rate. Section 5 (particularly the discussion at the end of Section 5.2) suggests that an improvement in learning rate can only occur if we make assumptions about $P(Y \mid X)$. Conversely, restricting the possible labeling functions may not lead to more than constant improvements. To see this, consider encoding the manifold assumption in the framework of Section 4.1, which immediately restricts the possible labelling functions, but [22] shows we can only get constant improvements in that case. The difference is that, in the framework of Section 4.1, one cannot infer enough about $P(Y|X)$. That is, even if we know that the best solution is one that separates two clusters, deciding which cluster belongs to which class can still be of worst-case order $\frac{1}{\sqrt{n}}$ by the same arguments as in Theorem 9 from [12].

7.6 SSL in Deep Learning

SSL has seen a resurgence in this era of deep learning, where, in some settings, significant performance improvements have been reported. The reader may wonder, therefore, why none of such works are covered in this survey. The reason is that we present strict mathematical analyses for possible improvements through SSL, something that remains elusive in the deep learning literature. Nevertheless, we want to sketch here what has been done at the intersection of deep learning and SSL and how this relates to topics covered in this survey.

The two paradigms that have been adopted for deep learning models are entropy and consistency regularization.

The main idea of entropy regularization [69], [70] is that we try to enforce low entropy predictions on the unlabeled data, which is equivalent to the decision boundary being in a low density region. In the deep learning community, this idea became known as *pseudo-labeling* [70] and is effectively a revival of self-learning, which was proposed in [71] as early as 1968. [72] shows that this procedure minimizes entropy and [73] demonstrates that this procedure may actually close a sample-complexity gap between standard and *robust classification*. The latter relies on a performance metric designed to study classification under adversarial attacks. As such, SLL may play a special role for deep models, which are known to be sensitive to adversarial attacks [74], [75].

Consistency regularization [76], [77], [78] exploits the idea that if we transform an unlabeled data point u in a meaningful way into $\hat{u}$, e.g., the slight rotation of an image, then the predictions $h(u)$ and $h(\hat{u})$ should be similar. The idea is thus to add a regularizer of the form $d(h(u), h(\hat{u}))$ to the loss term, where d is some sort of distance function. This idea actually ties directly in with the results presented in Section 4.1 and we suspect that similar performance guarantees hold. It would be of interest to see how the complexity of a neural network class shrinks under a consistency regularization method. In an optimistic mood, one may even hope that, in this way, one can generate non-vacuous performance guarantees with classical statistical learning theory.

7.7 Beyond PAC-Learnability

Arguably, the most general results of this survey are formulated in the PAC-learning framework, as presented in Sections 2.1.6 and 2.2. An exciting new type of learning framework was recently proposed in [44] and designated *universal learning*. The difference between this and PAC-learning is in essence the relationship between the error bound and the distribution P over $\mathcal{X} \times \mathcal{Y}$. In PAC-learnability, any error bound has to hold *uniformly* over all distributions, i.e., it is the *same* bound for all distributions. In universal learning, we can have constants in the bound that depend on the distribution. This has some dramatic consequences. In realizable PAC-learning, we either cannot learn at all or we learn with a linear rate. In universal learning, we have a trichotomy into linear, exponential, or arbitrarily slow rates. That this may play an important role in the interpretation of results can be seen through the analysis in Section 5.2. For each fixed distribution one may show an exponential learning rate, while a strict PAC-analysis leaves us with a slow rate, at least without further assumptions.

Acknowledgments

The authors thank Christina Göpfert wholeheartedly for the fruitful discussions that, among others, helped to identify the open questions in the field.

References

[1] S. Ben-David, T. Lu, and D. Pál, "Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 33–44.

[2] F. Cozman and I. Cohen, "Risks of semi-supervised learning: How unlabeled data can degrade performance of generative classifiers," in *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006, pp. 57–72.

[3] J. E. Van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Mach. Learn.*, vol. 109, no. 2, pp. 373–440, 2020.

[4] O. Chapelle, B. Schölkopf, and A. Zien, *Semi-Supervised Learning*, 1st ed. Cambridge, MA, USA: MIT Press, 2010.

[5] V. N. Vapnik, *Statistical Learning Theory*. New York, NY, USA: Wiley, 1998.

[6] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. Int. Conf. Mach. Learn.*, 1999, pp. 200–209.

[7] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold regularization: A geometric framework for learning from labeled and unlabeled examples," *J. Mach. Learn. Res.*, vol. 7, pp. 2399–2434, 2006.

[8] S. Shalev-Shwartz and S. Ben-David, *Understanding Machine Learning: From Theory to Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2014.

[9] M. Mohri, A. Rostamizadeh, and A. Talwalkar, *Foundations of Machine Learning*. Cambridge, MA, USA: MIT Press, 2012.

[10] J. Peters, D. Janzing, and B. Schölkopf, *Elements of Causal Inference: Foundations and Learning Algorithms*. Cambridge, MA, USA: MIT Press, 2017.

[11] M. Seeger, "Input-dependent regularization of conditional density models," Inst. ANC, Edinburgh, UK, Tech. Rep., 2000.

[12] C. Göpfert, S. Ben-David, O. Bousquet, S. Gelly, I. Tolstikhin, and R. Urner, "When can unlabeled data improve the learning rate?," in *Proc. Annu. Conf. Learn. Theory*, 2019, pp. 1500–1518.

[13] L. K. Hansen, "On Bayesian transduction: Implications for the covariate shift problem," in *Dataset Shift in Machine Learning*, Cambridge, U.K.: Cambridge, 2009, pp. 65–72.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *J. Roy. Statist. Soc.*, vol. 39, no. 1, pp. 1–38, 1977.

[15] T. Zhang and F. J. Oles, "A probability analysis on the value of unlabeled data for classification problems," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp.pp. 1191–1198.

[16] B. Schölkopf, D. Janzing, J. Peters, E. Sgouritsa, K. Zhang, and J. Mooij, "On causal and anticausal learning," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 1255–1262.

[17] M. Loog, "Contrastive pessimistic likelihood estimation for semi-supervised classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 462–475, Mar. 2016.

[18] J. H. Krijthe and M. Loog, "The pessimistic limits and possibilities of margin-based losses in semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2018, pp. 1795–1804.

[19] J. D. Lafferty and L. A. Wasserman, "Statistical analysis of semi-supervised regression," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 801–808.

[20] P. J. Bickel and B. Li, "Local polynomial regression on unknown manifolds," in *Complex Datasets and Inverse Problems*. Florence, Italy: Institute of Mathematical Statistics, 2007, pp. 177–186.

[21] M. Darnstädt, H. U. Simon, and B. Szörényi, "Unlabeled data does provably help," in *Proc. 30th Int. Symp. Theor. Aspects Comput. Sci.*, 2013, pp. 185–196.

[22] A. Mey, T. J. Viering, and M. Loog, "A distribution dependent and independent complexity analysis of manifold regularization," in *Proc. Int. Symp. Intell. Data Anal.*, 2020, pp. 326–338.

[23] A. Golovnev, D. Pál, and B. Szörényi, "The information-theoretic value of unlabeled data in semi-supervised learning," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 2328–2336.

[24] A. Globerson, R. Livni, and S. Shalev-Shwartz, "Effective semisupervised learning on manifolds," in *Proc. Annu. Conf. Learn. Theory*, 2017, pp. 978–1003.

[25] P. Niyogi, "Manifold regularization and semi-supervised learning: Some theoretical analyses," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1229–1250, 2013.

[26] N. Sokolovska, O. Cappé, and F. Yvon, "The asymptotics of semi-supervised learning in discriminative probabilistic models," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 984–991.

[27] M. Kääriäinen, "Generalization error bounds using unlabeled data," in *Proc. Annu. Conf. Learn. Theory*, 2005, pp. 127–142.

[28] M. Kawakita and T. Kanamori, "Semi-supervised learning with density-ratio estimation," *Mach. Learn.*, vol. 91, no. 2, pp. 189–209, 2013.

[29] A. B. Tsybakov, "Optimal aggregation of classifiers in statistical learning," *Ann. Statist.*, vol. 32, no. 1, pp. 135–166, 2004.

[30] D. A. McAllester, "Pac-Bayesian stochastic model selection," *Mach. Learn.*, vol. 51, no. 1, pp. 5–21, 2003.

[31] J. Langford and J. Shawe-Taylor, "Pac-bayes & margins," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2002, pp. 439–446.

[32] M.-F. Balcan and A. Blum, "A discriminative model for semi-supervised learning," *J. ACM*, vol. 57, no. 3, pp. 19:1–19:46, 2010.

[33] K. Sridharan and S. M. Kakade, "An information theoretic framework for multi-view learning," in *Proc. Annu. Conf. Learn. Theory*, 2008, pp. 403–414.

[34] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.

[35] A. Blum and T. Mitchell, "Combining labeled and unlabeled data with co-training," in *Proc. Annu. Conf. Learn. Theory*, 1998, pp. 92–100.

[36] A. Blum and S. Chawla, "Learning from labeled and unlabeled data using graph mincuts," in *Proc. Int. Conf. Mach. Learn.*, 2001, pp. 19–26.

[37] D. S. Rosenberg and P. L. Bartlett, "The Rademacher complexity of co-regularized kernel classes," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2007, pp. 396–403.

[38] J. Farquhar, D. Hardoon, H. Meng, J. Shawe-taylor, and S. Szedmák, "Two view learning: SVM-2k, theory and practice," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2006, pp. 355–362.

[39] V. Sindhwani and D. S. Rosenberg, "An RKHs for multi-view learning and manifold co-regularization," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 976–983.

[40] V. Castelli and T. M. Cover, "On the exponential value of labeled samples," *Pattern Recognit. Lett.*, vol. 16, no. 1, pp. 105–111, 1995.

[41] V. Castelli and T. M. Cover, "The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter," *IEEE Trans. Inf. Theory*, vol. 42, no. 6, pp. 2102–2117, Nov. 1996.

[42] K. Sinha and M. Belkin, "The value of labeled and unlabeled examples when the model is imperfect," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2007, pp. 1361–1368.

[43] P. Rigollet, "Generalization error bounds in semi-supervised classification under the cluster assumption," *J. Mach. Learn. Res.*, vol. 8, pp. 1369–1392, 2007.

[44] O. Bousquet, S. Hanneke, S. Moran, R. van Handel, and A. Yehudayoff, "A theory of universal learning," 2020, *arXiv: 2011.04483*.

[45] A. Singh, R. D. Nowak, and X. Zhu, "Unlabeled data: Now it helps, now it doesn't," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2008, pp. 1513–1520.

[46] M. Azizyan et al., "Density-sensitive semisupervised inference," *Ann. Statist.*, vol. 41, no. 2, pp. 751–771, 2013.

[47] D. Pechyony, "Theory and practice of transductive learning," Ph.D. dissertation, Isreal Inst. of Technol., Haifa, Israel, 2008.

[48] V. Vapnik, *Estimation of Dependences Based on Empirical Data*. Berlin, Germany: Springer-Verlag, 1982.

[49] P. Derbeko, R. El-Yaniv, and R. Meir, "Explicit learning curves for transduction and application to clustering and compression algorithms," *J. Artif. Intell. Res.*, vol. 22, pp. 117–142, 2004.

[50] D. McAllester, "Simplified pac-bayesian margin bounds," in *Proc. Learn. Theory Kernel Mach.*, 2003, pp. 203–215.

[51] D. A. McAllester, "Pac-bayesian stochastic model selection," *Mach. Learn.*, vol. 51, no. 1, pp. 5–21, 2003.

[52] G. M. Benedek and A. Itai, "Learnability with respect to fixed distributions," *Theor. Comput. Sci.*, vol. 86, no. 2, pp. 377–389, 1991.

[53] R. El-Yaniv and D. Pechyony, "Stable transductive learning," in *Proc. Annu. Conf. Learn. Theory*, 2006, pp. 35–49.

[54] O. Bousquet and A. Elisseeff, "Stability and generalization," *J. Mach. Learn. Res.*, vol. 2, pp. 499–526, 2002.

[55] S. Kutin and P. Niyogi, "Almost-everywhere algorithmic stability and generalization error," in *Proc. Conf. Uncertainty Artif. Intell.*, 2002, pp. 275–282.

[56] S. Kutin, "Extensions to Mcdiarmid's inequality when differences are bounded with high probability," Dep. Comput. Sci., Univ. Chicago, Tech. Rep. TR-2002–04, 2002.

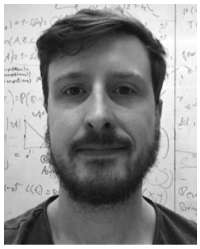
[57] R. Johnson and T. Zhang, "On the effectiveness of laplacian normalization for graph semi-supervised learning," *J. Mach. Learn. Res.*, vol. 8, pp. 1489–1517, 2007.

[58] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Ann. Statist.*, vol. 33, no. 4, pp. 1497–1537, 2005.

[59] R. El-Yaniv and D. Pechyony, "Transductive Rademacher complexity and its applications," *J. Artif. Intell. Res.*, vol. 35, no. 1, pp. 193–234, 2009.

[60] Y. Maximov, M.-R. Amini, and Z. Harchaoui, "Rademacher complexity bounds for a penalized multiclass semi-supervised algorithm," in *Proc. Int. Joint Conf. Artif. Intell.*, 2018, pp. 5637–5641.

[61] P. L. Bartlett and S. Mendelson, "Rademacher and Gaussian complexities: Risk bounds and structural results," *J. Mach. Learn. Res.*, vol. 3, pp. 463–482, 2003.
[62] G. R. G. Lanckriet, N. Cristianini, P. L. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *J. Mach. Learn. Res.*, vol. 5, pp. 27–72, 2004.
[63] X. Zhu, J. Kandola, J. Lafferty, and Z. Ghahramani, "Graph kernels by spectral transforms," in *Semi-Supervised Learning*. Cambridge, MA, USA: MIT Press, 2006, pp. 277–291.
[64] R. Johnson and T. Zhang, "Graph-based semi-supervised learning and spectral kernel design," *IEEE Trans. IT*, vol. 54, no. 1, pp. 275–288, Jan. 2008.
[65] Y.-F. Li and Z.-H. Zhou, "Towards making unlabeled data never hurt," in *Proc. Int. Conf. Mach. Learn.*, 2011, pp. 1081–1088.
[66] K. P. Bennett and A. Demiriz, "Semi-supervised support vector machines," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 1999, pp. 368–374.
[67] J. H. Krijthe and M. Loog, "Projected estimators for robust semi-supervised classification," *Mach. Learn.*, vol. 106, no. 7, pp. 993–1008, 2017.
[68] M.-F. Balcan, E. Blais, A. Blum, and L. Yang, "Active property testing," in *Proc. IEEE 53rd Annu. Symp. Found. Comput. Sci.*, 2012, pp. 21–30.
[69] Y. Grandvalet and Y. Bengio, "Semi-supervised learning by entropy minimization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2004, pp. 529–536.
[70] D.-H. Lee, "Pseudo-label : The simple and efficient semi-supervised learning method for deep neural networks," in *Proc. Int. Conf. Mach. Learn. Workshop Track*, 2013, Art. no. 896.
[71] H. O. Hartley and J. N. Rao, "Classification and estimation in analysis of variance problems," *Revue de l'Institut Int. de Statistique*, vol. 36, no. 2, pp. 141–147, 1968.
[72] S. P. Abney, "Understanding the Yarowsky algorithm," *Comput. Linguistics*, vol. 30, no. 3, pp. 365–395, 2004.
[73] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. Liang, "Unlabeled data improves adversarial robustness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, pp. 11190–11201.
[74] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.
[75] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Representations Workshop Track*, 2017.
[76] P. Bachman, O. Alsharif, and D. Precup, "Learning with pseudo-ensembles," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3365–3373.
[77] M. Sajjadi, M. Javanmardi, and T. Tasdizen, "Regularization with stochastic transformations and perturbations for deep semi-supervised learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 1171–1179.
[78] S. Laine and T. Aila, "Temporal ensembling for semi-supervised learning," in *Proc. Int. Conf. Learn. Representations*, 2017.



Alexander Mey received the MSc degree in mathematics from University Bonn and the PhD degree from Pattern Recognition Laboratory, TU Delft. He is currently working as a postdoc with the Interactive Intelligence Group at TU Delft. His research interests lie within the theoretical foundations of all types of intelligent systems.

Marco Loog received the MSc degree in mathematics from Utrecht University and the PhD degree from the Image Sciences Institute. He worked as a scientist with the IT University of Copenhagen, the University of Copenhagen, and Nordic Bioscience and now is with Delft University of Technology to research and to teach. He is a honorary professor with the University of Copenhagen. His principal research interest is with supervised learning in all sorts of shapes and sizes.