A Distribution Dependent and Independent Complexity Analysis of Manifold Regularization

Mey, Alexander; Viering, Tom Julian; Loog, Marco

DOI
10.1007/978-3-030-44584-3_26

Publication date
2020

Document Version
Final published version

Published in
Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Proceedings

Citation (APA)
Mey, A., Viering, T. J., & Loog, M. (2020). A Distribution Dependent and Independent Complexity Analysis of Manifold Regularization. In M. R. Berthold, A. Feelders, & G. Krempl (Eds.), *Advances in Intelligent Data Analysis XVIII - 18th International Symposium on Intelligent Data Analysis, IDA 2020, Proceedings* (Vol. 12080, pp. 326-338). (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics); Vol. 12080 LNCS). SpringerOpen. https://doi.org/10.1007/978-3-030-44584-3_26

A Distribution Dependent and Independent Complexity Analysis of Manifold Regularization

Alexander Mey[1(✉)], Tom Julian Viering[1], and Marco Loog[1,2]

[1] Delft University of Technology, Delft, The Netherlands
{a.mey,t.j.viering}@tudelft.nl
[2] University of Copenhagen, Copenhagen, Denmark
m.loog@tudelft.nl

Abstract. Manifold regularization is a commonly used technique in semi-supervised learning. It enforces the classification rule to be smooth with respect to the data-manifold. Here, we derive sample complexity bounds based on pseudo-dimension for models that add a convex data dependent regularization term to a supervised learning process, as is in particular done in Manifold regularization. We then compare the bound for those semi-supervised methods to purely supervised methods, and discuss a setting in which the semi-supervised method can only have a constant improvement, ignoring logarithmic terms. By viewing Manifold regularization as a kernel method we then derive Rademacher bounds which allow for a distribution *dependent* analysis. Finally we illustrate that these bounds may be useful for choosing an appropriate manifold regularization parameter in situations with very sparsely labeled data.

Keywords: Semi-supervised learning · Learning theory · Manifold regularization

1 Introduction

In many applications, as for example image or text classification, gathering unlabeled data is easier than gathering labeled data. Semi-supervised methods try to extract information from the unlabeled data to get improved classification results over purely supervised methods. A well-known technique to incorporate unlabeled data into a learning process is manifold regularization (MR) [7,18]. This procedure adds a data-dependent penalty term to the loss function that penalizes classification rules that behave non-smooth with respect to the data distribution. This paper presents a sample complexity and a Rademacher complexity analysis for this procedure. In addition it illustrates how our Rademacher complexity bounds may be used for choosing a suitable Manifold regularization parameter.

We organize this paper as follows. In Sects. 2 and 3 we discuss related work and introduce the semi-supervised setting. In Sect. 4 we formalize the idea of

M. R. Berthold et al. (Eds.): IDA 2020, LNCS 12080, pp. 326–338, 2020.
https://doi.org/10.1007/978-3-030-44584-3_26

adding a distribution-dependent penalty term to a loss function. Algorithms such as manifold, entropy or co-regularization [7,14,21] follow this idea. Section 5 generalizes a bound from [4] to derive sample complexity bounds for the proposed framework, and thus in particular for MR. For the specific case of regression, we furthermore adapt a sample complexity bound from [1], which is essentially tighter than the first bound, to the semi-supervised case. In the same section we sketch a setting in which we show that if our hypothesis set has finite pseudo-dimension, and we ignore logarithmic factors, any semi-supervised learner (SSL) that falls in our framework has at most a constant improvement in terms of sample complexity. In Sect. 6 we show how one can obtain distribution *dependent* complexity bounds for MR. We review a kernel formulation of MR [20] and show how this can be used to estimate Rademacher complexities for *specific* datasets. In Sect. 7 we illustrate on an artificial dataset how the distribution dependent bounds could be used for choosing the regularization parameter of MR. This is particularly useful as the analysis does not need an additional labeled validation set. The practicality of this approach requires further empirical investigation. In Sect. 8 we discuss our results and speculate about possible extensions.

2 Related Work

In [13] we find an investigation of a setting where distributions on the input space $\mathcal{X}$ are restricted to ones that correspond to unions of irreducible algebraic sets of a fixed size $k \in \mathbb{N}$, and each algebraic set is either labeled 0 or 1. A SSL that knows the true distribution on $\mathcal{X}$ can identify the algebraic sets and reduce the hypothesis space to all 2^k possible label combinations on those sets. As we are left with finitely many hypotheses we can learn them efficiently, while they show that every supervised learner is left with a hypothesis space of infinite VC dimension.

The work in [18] considers manifolds that arise as embeddings from a circle, where the labeling over the circle is (up to the decision boundary) smooth. They then show that a learner that has knowledge of the manifold can learn efficiently while for every fully supervised learner one can find an embedding and a distribution for which this is not possible.

The relation to our paper is as follows. They provide specific examples where the sample complexity between a semi-supervised and a supervised learner are infinitely large, while we explore general sample complexity bounds of MR and sketch a setting in which MR can not essentially improve over supervised methods.

3 The Semi-supervised Setting

We work in the statistical learning framework: we assume we are given a feature domain $\mathcal{X}$ and an output space $\mathcal{Y}$ together with an unknown probability distribution P over $\mathcal{X} \times \mathcal{Y}$. In binary classification we usually have that $\mathcal{Y} = \{-1, 1\}$, while for regression $\mathcal{Y} = \mathbb{R}$. We use a loss function $\phi : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, which is convex in the first argument and in practice usually a surrogate for the 0–1 loss

in classification, and the squared loss in regression tasks. A hypothesis f is a function $f : \mathcal{X} \to \mathbb{R}$. We set (X, Y) to be a random variable distributed according to P, while small x and y are elements of $\mathcal{X}$ and $\mathcal{Y}$ respectively. Our goal is to find a hypothesis f, within a restricted class $\mathcal{F}$, such that the expected loss $Q(f) := \mathbb{E}[\phi(f(X), Y)]$ is small. In the standard supervised setting we choose a hypothesis f based on an i.i.d. sample $S_n = \{(x_i, y_i)\}_{i\in\{1,..,n\}}$ drawn from P. With that we define the empirical risk of a model $f \in \mathcal{F}$ with respect to ϕ and measured on the sample S_n as $\hat{Q}(f, S_n) = \frac{1}{n}\sum_{i=1}^{n} \phi(f(x_i), y_i)$. For ease of notation we sometimes omit S_n and just write $\hat{Q}(f)$. Given a learning problem defined by $(P, \mathcal{F}, \phi)$ and a labeled sample S_n, one way to choose a hypothesis is by the empirical risk minimization principle

$$f_{\text{sup}} = \arg\min_{f\in\mathcal{F}} \hat{Q}(f, S_n). \tag{1}$$

We refer to f_{sup} as the *supervised solution*. In SSL we additionally have samples with unknown labels. So we assume to have $n + m$ samples $(x_i, y_i)_{i\in\{1,..,n+m\}}$ independently drawn according to P, where y_i has not been observed for the last m samples. We furthermore set $U = \{x_1, ..., x_{x_n+m}\}$, so U is the set that contains all our available information about the feature distribution.

Finally we denote by $m^L(\epsilon, \delta)$ the sample complexity of an algorithm L. That means that for all $n \geq m^L(\epsilon, \delta)$ and all possible distributions P the following holds. If L outputs a hypothesis f_L after seeing an n-sample, we have with probability of at least $1 - \delta$ over the n-sample S_n that $Q(f_L) - \min_{f\in\mathcal{F}} Q(f) \leq \epsilon$.

4 A Framework for Semi-supervised Learning

We follow the work of [4] and introduce a second convex loss function $\psi : \mathcal{F}\times\mathcal{X} \to \mathbb{R}_+$ that only depends on the input feature and a hypothesis. We refer to ψ as the *unsupervised loss* as it does not depend on any labels. We propose to *add* the unlabeled data through the loss function ψ and add it as a penalty term to the supervised loss to obtain the semi-supervised solution

$$f_{\text{semi}} = \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \phi(f(x_i), y_i) + \lambda \frac{1}{n+m}\sum_{j=1}^{n+m} \psi(f, x_j), \tag{2}$$

where $\lambda > 0$ controls the trade-off between the supervised and the unsupervised loss. This is in contrast to [4], as they use the unsupervised loss to restrict the hypothesis space directly. In the following section we recall the important insight that those two formulations are equivalent in some scenarios and we can use [4] to generate sample complexity bounds for the here presented SSL framework.

For ease of notation we set $\hat{R}(f, U) = \frac{1}{n+m}\sum_{j=1}^{n+m} \psi(f, x_j)$ and $R(f) = \mathbb{E}[\psi(f, X)]$. We do not claim any novelty for the idea of adding an unsupervised loss for regularization. A different framework can be found in [11, Chapter 10]. We are, however, not aware of a deeper analysis of this particular formulation, as done for example by the sample complexity analysis in this paper. As we are in particular interested in the class of MR schemes we first show that this method fits our framework.

Example: Manifold Regularization. Overloading the notation we write now $P(X)$ for the distribution P restricted to $\mathcal{X}$. In MR one assumes that the input distribution $P(X)$ has support on a compact manifold $M \subset \mathcal{X}$ and that the predictor $f \in \mathcal{F}$ varies smoothly in the geometry of M [7]. There are several regularization terms that can enforce this smoothness, one of which is $\int_M ||\nabla_M f(x)||^2 dP(x)$, where $\nabla_M f$ is the gradient of f along M. We know that $\int_M ||\nabla_M f(x)||^2 dP(x)$ may be approximated with a finite sample of $\mathcal{X}$ drawn from $P(X)$ [6]. Given such a sample $U = \{x_1, ..., x_{n+m}\}$ one defines first a weight matrix W, where $W_{ij} = e^{-||x_i - x_j||^2/\sigma}$. We set L then as the Laplacian matrix $L = D - W$, where D is a diagonal matrix with $D_{ii} = \sum_{j=1}^{n+m} W_{ij}$. Let furthermore $f_U = (f(x_1), ..., f(x_{n+m}))^t$ be the evaluation vector of f on U. The expression $\frac{1}{(n+m)^2} f_U^t L f_U = \frac{1}{(n+m)^2} \sum_{i,j} (f(x_i) - f(x_j))^2 W_{ij}$ converges to $\int_M ||\nabla_M f||^2 dP(x)$ under certain conditions [6]. This motivates us to set the unsupervised loss as $\psi(f, (x_i, x_j)) = (f(x_i) - f(x_j))^2 W_{ij}$. Note that $f_U^t L f_U$ is indeed a convex function in f: As L is a Laplacian matrix it is positive definite and thus $f_U^t L f_U$ defines a norm in f. Convexity follows then from the triangle inequality.

5 Analysis of the Framework

In this section we analyze the properties of the solution f_{semi} found in Equation (2). We derive sample complexity bounds for this procedure, using results from [4], and compare them to sample complexities for the supervised case. In [4] the unsupervised loss is used to restrict the hypothesis space directly, while we use it as a regularization term in the empirical risk minimization as usually done in practice. To switch between the views of a constrained optimization formulation and our formulation (2) we use the following classical result from convex optimization [15, Theorem 1].

Lemma 1. *Let $\phi(f(x), y)$ and $\psi(f, x)$ be functions convex in f for all x, y. Then the following two optimization problems are equivalent:*

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(f(x_i), y_i) + \lambda \frac{1}{n+m} \sum_{i=1}^{n+m} \psi(f, x_i) \tag{3}$$

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \phi(f(x_i), y_i) \quad \textit{subject to} \quad \sum_{i=1}^{n+m} \frac{1}{n+m} \psi(f, x_i) \leq \tau \tag{4}$$

Where equivalence means that for each λ we can find a τ such that both problems have the same solution and vice versa.

For our later results we will need the conditions of this lemma are true, which we believe to be not a strong restriction. In our sample complexity analysis we stick as close as possible to the actual formulation and implementation of MR, which is usually a convex optimization problem. We first turn to our sample complexity bounds.

5.1 Sample Complexity Bounds

Sample complexity bounds for supervised learning use typically a notion of complexity of the hypothesis space to bound the worst case difference between the estimated and the true risk. As our hypothesis class allows for real-valued functions, we will use the notion of pseudo-dimension $\mathrm{Pdim}(\mathcal{F}, \phi)$, an extension of the VC-dimension to real valued loss functions ϕ and hypotheses classes $\mathcal{F}$ [17,22]. Informally speaking, the pseudo-dimension is the VC-dimension of the set of functions that arise when we threshold real-valued functions to define binary functions. Note that sometimes the pseudo-dimension will have as input the loss function, and sometimes not. This is because some results use the concatenation of loss function and hypotheses to determine the capacity, while others only use the hypotheses class. This lets us state our first main result, which is a generalization of [4, Theorem 10] to bounded loss functions and real valued function spaces.

Theorem 1. *Let $\mathcal{F}_\tau^\psi := \{f \in \mathcal{F} \mid \mathbb{E}[\psi(f, x)] \leq \tau\}$. Assume that ϕ, ψ are measurable loss functions such that there exists constants $B_1, B_2 > 0$ with $\psi(f, x) \leq B_1$ and $\phi(f(x), y) \leq B_2$ for all x, y and $f \in \mathcal{F}$ and let P be a distribution. Furthermore let $f_\tau^* = \arg\min_{f \in \mathcal{F}_\tau^\psi} Q(f)$. Then an unlabeled sample U of size*

$$m \geq \frac{8{B_1}^2}{\epsilon^2}\left[\ln\frac{16}{\delta} + 2\,\mathrm{Pdim}(\mathcal{F}, \psi)\ln\frac{4B_1}{\epsilon} + 1\right] \tag{5}$$

and a labeled sample S_n of size

$$n \geq \max\left(\frac{8{B_2}^2}{\epsilon^2}\left[\ln\frac{8}{\delta} + 2\,\mathrm{Pdim}(\mathcal{F}^\psi_{\tau+\frac{\epsilon}{2}}, \phi)\ln\frac{4B_2}{\epsilon} + 1\right], \frac{h}{4}\right) \tag{6}$$

is sufficient to ensure that with probability at least $1-\delta$ the classifier $g \in \mathcal{F}$ that minimizes $\hat{Q}(\cdot, S_n)$ subject to $\hat{R}(\cdot, U) \leq \tau + \frac{\epsilon}{2}$ satisfies

$$Q(g) \leq Q(f_\tau^*) + \epsilon. \tag{7}$$

Sketch Proof: The idea is to combine three partial results with a union bound. For the first part we use Theorem 5.1 from [22] with $h = \mathrm{Pdim}(\mathcal{F}, \psi)$ to show that an unlabeled sample size of

$$m \geq \frac{8{B_1}^2}{\epsilon^2}\left[\ln\frac{16}{\delta} + 2h\ln\frac{4B_1}{\epsilon} + 1\right] \tag{8}$$

is sufficient to guarantee $\hat{R}(f) - R(f) < \frac{\epsilon}{2}$ for all $f \in \mathcal{F}$ with probability at least $1 - \frac{\delta}{4}$. In particular choosing $f = f_\tau^*$ and noting that by definition $R(f_\tau^*) \leq \tau$ we conclude that with the same probability

$$\hat{R}(f_\tau^*) \leq \tau + \frac{\epsilon}{2}. \tag{9}$$

For the second part we use Hoeffding's inequality to show that the labeled sample size is big enough that with probability at least $1 - \frac{\delta}{4}$ it holds that

$$\hat{Q}(f_\tau^*) \leq Q(f_\tau^*) + B_2\sqrt{\ln(\frac{4}{\delta})\frac{1}{2n}}. \tag{10}$$

The third part again uses Th. 5.1 from [22] with $h = \text{Pdim}(\mathcal{F}_\tau^\psi, \phi)$ to show that $n \geq \frac{8{B_2}^2}{\epsilon^2}\left[\ln\frac{8}{\delta} + 2h\ln\frac{4B_2}{\epsilon} + 1\right]$ is sufficient to guarantee $Q(f) \leq \hat{Q}(f) + \frac{\epsilon}{2}$ with probability at least $1 - \frac{\delta}{2}$.

Putting everything together with the union bound we get that with probability $1 - \delta$ the classifier g that minimizes $\hat{Q}(\cdot, X, Y)$ subject to $\hat{R}(\cdot, U) \leq \tau + \frac{\epsilon}{2}$ satisfies

$$Q(g) \leq \hat{Q}(g) + \frac{\epsilon}{2} \leq \hat{Q}(f_\tau^*) + \frac{\epsilon}{2} \leq Q(f_\tau^*) + \frac{\epsilon}{2} + B_2\sqrt{\frac{\ln(\frac{4}{\delta})}{2n}}. \tag{11}$$

Finally the labeled sample size is big enough to bound the last rhs term by $\frac{\epsilon}{2}$. □

The next subsection uses this theorem to derive sample complexity bounds for MR. First, however, a remark about the assumption that the loss function ϕ is globally bounded. If we assume that $\mathcal{F}$ is a reproducing kernel Hilbert space there exists an $M > 0$ such that for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$ it holds that $|f(x)| \leq M||f||_{\mathcal{F}}$. If we restrict the norm of f by introducing a regularization term with respect to the norm $||.||_{\mathcal{F}}$, we know that the image of $\mathcal{F}$ is globally bounded. If the image is also closed it will be compact, and thus ϕ will be globally bounded in many cases, as most loss functions are continuous. This can also be seen as a justification to also use an intrinsic regularization for the norm of f in addition to the regularization by the unsupervised loss, as only then the guarantees of Theorem 1 apply. Using this bound together with Lemma 1 we can state the following corollary to give a PAC-style guarantee for our proposed framework.

Corollary 1. *Let ϕ and ψ be convex supervised and an unsupervised loss function that fulfill the assumptions of Theorem 1. Then f_{semi} (2) satisfies the guarantees given in Theorem 1, when we replace for it g in Inequality (7).*

Recall that in the MR setting $\hat{R}(f) = \frac{1}{(n+m)^2}\sum_{i=1}^{n+m} W_{ij}(f(x_i) - f(x_j))^2$. So we gather unlabeled samples from $\mathcal{X} \times \mathcal{X}$ instead of $\mathcal{X}$. Collecting m samples from $\mathcal{X}$ equates $m^2 - 1$ samples from $\mathcal{X} \times \mathcal{X}$ and thus we only need $\sqrt{m}$ instead of m unlabeled samples for the same bound.

5.2 Comparison to the Supervised Solution

In the SSL community it is well-known that using SSL does not come without a risk [11, Chapter 4]. Thus it is of particular interest how those methods compare to purely supervised schemes. There are, however, many potential supervised methods we can think of. In many works this problem is avoided by comparing

to all possible supervised schemes [8,12,13]. The framework introduced in this paper allows for a more fine-grained analysis as the semi-supervision happens on top of an already existing supervised methods. Thus, for our framework, it is natural to compare the sample complexities of f_{sup} with the sample complexity of f_{semi}. To compare the supervised and semi-supervised solution we will restrict ourselves to the square loss. This allows us to draw from [1, Chapter 20], where one can find lower and upper sample complexity bounds for the regression setting. The main insight from [1, Chapter 20] is that the sample complexity depends in this setting on whether the hypothesis class is (closure) convex or not. As we anyway need convexity of the space, which is stronger than closure convexity, to use Lemma 1, we can adapt Theorem 20.7 from [1] to our semi-supervised setting.

Theorem 2. *Assume that $\mathcal{F}^{\psi}_{\tau+\epsilon}$ is a closure convex class with functions mapping to $[0,1]$[1], that $\psi(f,x) \leq B_1$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$ and that $\phi(f(x),y) = (f(x)-y)^2$. Assume further that there is a $B_2 > 0$ such that $(f(x)-y)^2 < B_2$ almost surely for all $(x,y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}^{\psi}_{\tau+\epsilon}$. Then an unlabeled sample size of*

$$m \geq \frac{2{B_1}^2}{\epsilon^2}\left[\ln\frac{8}{\delta} + 2\,\mathrm{Pdim}(\mathcal{F},\psi)\ln\frac{2B_1}{\epsilon} + 2\right] \tag{12}$$

and a labeled sample size of

$$n \geq \mathcal{O}\left(\frac{B_2^2}{\epsilon}\left(\mathrm{Pdim}(\mathcal{F}^{\psi}_{\tau+\epsilon})\ln\frac{\sqrt{B_2}}{\epsilon} + \ln\frac{2}{\delta}\right)\right) \tag{13}$$

is sufficient *to guarantee that with probability at least $1-\delta$ the classifier g that minimizes $\hat{Q}(\cdot)$ w.r.t $\hat{R}(f) \leq \tau + \epsilon$ satisfies*

$$Q(g) \leq \min_{f \in \mathcal{F}^{\psi}_{\tau}} Q(f) + \epsilon. \tag{14}$$

Proof: As in the proof of Theorem 1 the unlabeled sample size is sufficient to guarantee with probability at least $1-\frac{\delta}{2}$ that $R(f^*_{\tau}) \leq \tau+\epsilon$. The labeled sample size is big enough to guarantee with at least $1-\frac{\delta}{2}$ that $Q(g) \leq Q(f^*_{\tau+\epsilon}) + \epsilon$ [1, Theorem 20.7]. Using the union bound we have with probability of at least $1-\delta$ that $Q(g) \leq Q(f^*_{\tau+\epsilon}) + \epsilon \leq Q(f^*_{\tau}) + \epsilon$. □

Note that the previous theorem of course implies the same learning rate in the supervised case, as the only difference will be the pseudo-dimension term. As in specific scenarios this is also the best possible learning rate, we obtain the following negative result for SSL.

Corollary 2. *Assume that ϕ is the square loss, $\mathcal{F}$ maps to the interval $[0,1]$ and $\mathcal{Y} = [1-B, B]$ for a $B \geq 2$. If $\mathcal{F}$ and $\mathcal{F}^{\psi}_{\tau}$ are both closure convex, then for sufficiently small $\epsilon, \delta > 0$ it holds that $m^{sup}(\epsilon,\delta) = \tilde{\mathcal{O}}(m^{semi}(\epsilon,\delta))$, where*

[1] In the remarks after Theorem 1 we argue that in many cases |f(x)| is bounded, and in those cases we can always map to [0,1] by re-scaling.

$\tilde{\mathcal{O}}$ suppresses logarithmic factors, and m^{semi}, m^{sup} denote the sample complexity of the semi-supervised and the supervised learner respectively. In other words, the semi-supervised method can improve the learning rate by at most a constant which may depend on the pseudo-dimensions, ignoring logarithmic factors. Note that this holds in particular for the manifold regularization algorithm.

Proof: The assumptions made in the theorem allow is to invoke Equation (19.5) from [1] which states that $m^{\text{semi}} = \Omega(\frac{1}{\epsilon} + \text{Pdim}(\mathcal{F}_\tau^\psi))$.[2] Using Inequality (13) as an upper bound for the supervised method and comparing this to Eq. (19.5) from [1] we observe that all differences are either constant or logarithmic in ϵ and δ. □

5.3 The Limits of Manifold Regularization

We now relate our result to the conjectures published in [19]: A SSL cannot learn faster by more than a constant (which may depend on the hypothesis class $\mathcal{F}$ and the loss ϕ) than the supervised learner. Theorem 1 from [12] showed that this conjecture is true up to a logarithmic factor, much like our result, for classes with finite VC-dimension, and SSL that do *not* make any distributional assumptions. Corollary 2 shows that this statement also holds in some scenarios for all SSL that fall in our proposed framework. This is somewhat surprising, as our result holds explicitly for SSLs that *do* make assumptions about the distribution: MR assumes the labeling function behaves smoothly w.r.t. the underlying manifold.

6 Rademacher Complexity of Manifold Regularization

In order to find out in which scenarios semi-supervised learning can help it is useful to also look at distribution *dependent* complexity measures. For this we derive computational feasible upper and lower bounds on the Rademacher complexity of MR. We first review the work of [20]: they create a kernel such that the inner product in the corresponding kernel Hilbert space contains automatically the regularization term from MR. Having this kernel we can use standard upper and lower bounds of the Rademacher complexity for RKHS, as found for example in [10]. The analysis is thus similar to [21]. They consider a co-regularization setting. In particular [20, p. 1] show the following, here informally stated, theorem.

Theorem 3 ([20, Propositions 2.1, 2.2]). *Let H be a RKHS with inner product $\langle\cdot,\cdot\rangle_H$. Let $U = \{x_1, ..., x_{n+m}\}$, $f, g \in H$ and $f_U = (f(x_1), ..., f(x_{n+m}))^t$. Furthermore let $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ be any inner product in $\mathbb{R}^n$. Let $\tilde{H}$ be the same space of functions as H, but with a newly defined inner product by $\langle f, g\rangle_{\tilde{H}} = \langle f, g\rangle_H + \langle f_U, g_U\rangle_{\mathbb{R}^n}$. Then $\tilde{H}$ is a RKHS.*

[2] Note that the original formulation is in terms of the fat-shattering dimension, but this is always bounded by the pseudo-dimension.

Assume now that L is a positive definite n-dimensional matrix and we set the inner product $\langle f_U, g_U \rangle_{\mathbb{R}^n} = f_U^t L g_U$. By setting L as the Laplacian matrix (Sect. 4) we note that the norm of H automatically regularizes w.r.t. the data manifold given by $\{x_1, ..., x_{n+m}\}$. We furthermore know the exact form of the kernel of $\tilde{H}$.

Theorem 4 ([20, Proposition 2.2]). *Let $k(x, y)$ be the kernel of H, K be the gram matrix given by $K_{ij} = k(x_i, x_j)$ and $k_x = (k(x_1, x), ..., k(x_{n+m}, x))^t$. Finally let I be the $n + m$ dimensional identity matrix. The kernel of $\tilde{H}$ is then given by $\tilde{k}(x, y) = k(x, y) - k_x^t (I + LK)^{-1} L k_y$.*

This interpretation of MR is useful to derive computationally feasible upper and lower bounds of the empirical Rademacher complexity, giving distribution *dependent* complexity bounds. With $\sigma = (\sigma_1, ..., \sigma_n)$ i.i.d Rademacher random variables (i.e. $P(\sigma_i = 1) = P(\sigma_i = -1) = \frac{1}{2}$.), recall that the empirical Rademacher complexity of the hypothesis class H and measured on the sample labeled input features $\{x_1, ..., x_n\}$ is defined as

$$\mathrm{Rad}_n(H) = \frac{1}{n} \mathbb{E}_\sigma \sup_{f \in H} \sum_{i=1}^{n} \sigma_i f(x_i). \tag{15}$$

Theorem 5 ([10, p. 333]). *Let H be a RKHS with kernel k and $H_r = \{f \in H \mid ||f||_H \leq r\}$. Given an n sample $\{x_1, ..., x_n\}$ we can bound the empirical Rademacher complexity of H_r by*

$$\frac{r}{n\sqrt{2}} \sqrt{\sum_{i=1}^{n} k(x_i, x_i)} \leq \mathrm{Rad}_n(H_r) \leq \frac{r}{n} \sqrt{\sum_{i=1}^{n} k(x_i, x_i)}. \tag{16}$$

The previous two theorems lead to upper bounds on the complexity of MR, in particular we can bound the maximal reduction over supervised learning.

Corollary 3. *Let H be a RKHS and for $f, g \in H$ define the inner product $\langle f, g \rangle_{\tilde{H}} = \langle f, g \rangle_H + f_U(\mu L) g_U^t$, where L is a positive definite matrix and $\mu \in \mathbb{R}$ is a regularization parameter. Let $\tilde{H}_r$ be defined as before, then*

$$\mathrm{Rad}_n(\tilde{H}_r) \leq \frac{r}{n} \sqrt{\sum_{i=1}^{n} k(x_i, x_i) - k_{x_i}^t (\frac{1}{\mu} I + LK)^{-1} L k_{x_i}}. \tag{17}$$

Similarly we can obtain a lower bound in line with Inequality (16).

The corollary shows in particular that the difference of the Rademacher complexity of the supervised and the semi-supervised method is given by the term

$k^t_{x_i}(\frac{1}{\mu}I_{n+m} + LK)^{-1}Lk_{x_i}$. This can be used for example to compute generalization bounds [17, Chapter 3]. We can also use the kernel to compute local Rademacher complexities which may yield tighter generalization bounds [5]. Here we illustrate the use of our bounds for choosing the regularization parameter μ without the need for an additional labeled validation set.

7 Experiment: Concentric Circles

We illustrate the use of Eq. (17) for model selection. In particular, it can be used to get an initial idea of how to choose the regularization parameter μ. The idea is to plot the Rademacher complexity versus the parameter μ as in Fig. 1. We propose to use an heuristic which is often used in clustering, the so called elbow criteria [9]. We essentially want to find a μ such that increasing the μ will not result in much reduction of the complexity anymore. We test this idea on a dataset which consists out of two concentric circles with 500 datapoints in $\mathbb{R}^2$, 250 per circle, see also Fig. 2. We use a Gaussian base kernel with bandwidth set to 0.5. The MR matrix L is the Laplacian matrix, where weights are computed with a Gaussian kernel with bandwidth 0.2. Note that those parameters have to be carefully set in order to capture the structure of the dataset, but this is not the current concern: we assume we already found a reasonable choice for those parameters. We add a small L2-regularization that ensures that the radius r in Inequality (17) is finite. The precise value of r plays a secondary role as the behavior of the curve from Fig. 1 remains the same.

Looking at Fig. 1 we observe that for μ smaller than 0.1 the curve still drops steeply, while after 0.2 it starts to flatten out. We thus plot the resulting kernels for $\mu = 0.02$ and $\mu = 0.2$ in Fig. 2. We plot the isolines of the kernel around the point of class one, the red dot in the figure. We indeed observe that for $\mu = 0.02$ we don't capture that much structure yet, while for $\mu = 0.2$ the two concentric circles are almost completely separated by the kernel. If this procedure indeed elevates to a practical method needs further empirical testing.

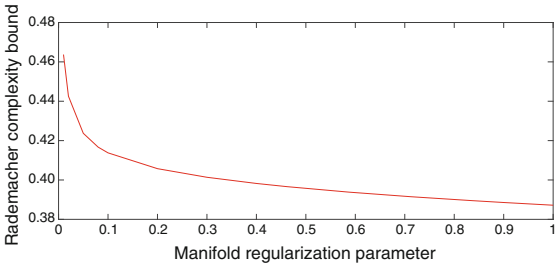
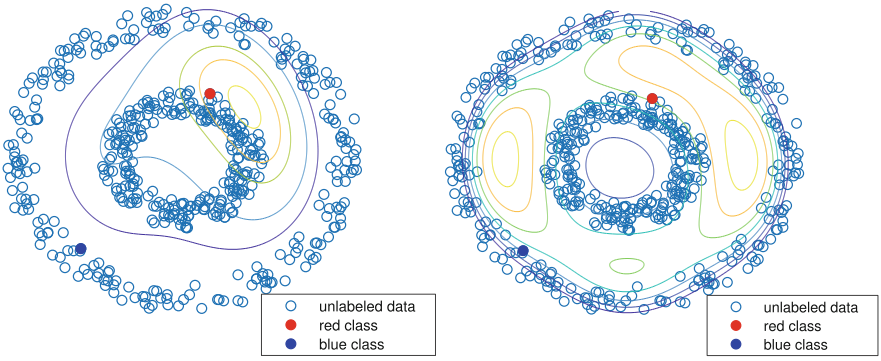


Fig. 1. The behavior of the Rademacher complexity when using manifold regularization on circle dataset with different regularization values μ.



(a) $\mu = 0.02$ (b) $\mu = 0.2$

Fig. 2. The resulting kernel when we use manifold regularization with parameter μ set to 0.02 and 0.2.

8 Discussion and Conclusion

This paper analysed improvements in terms of sample or Rademacher complexity for a certain class of SSL. The performance of such methods depends both on how the approximation error of the class $\mathcal{F}$ compares to that of $\mathcal{F}_\tau^\psi$ and on the reduction of complexity by switching from the first to the latter. In our analysis we discussed the second part. The first part depends on a notion the literature often refers to as a *semi-supervised assumption*. This assumption basically states that we can learn with $\mathcal{F}_\tau^\psi$ as good as with $\mathcal{F}$. Without prior knowledge, it is unclear whether one can test efficiently if the assumption is true or not. Or is it possible to treat just this as a model selection problem? The only two works we know that provide some analysis in this direction are [3], which discusses the sample consumption to test the so-called cluster assumption, and [2], which analyzes the overhead of cross-validating the hyper-parameter coming from their proposed semi-supervised approach.

As some of our settings need restrictions, it is natural to ask whether we can extend the results. First, Lemma 1 restricts us to convex optimization problems. If that assumption would be unnecessary, one may get interesting extensions. Neural networks, for example, are typically not convex in their function space and we cannot guarantee the fast learning rate from Theorem 2. But maybe there are semi-supervised methods that turn this space convex, and thus could achieve fast rates. In Theorem 2 we have to restrict the loss to be the square loss, and [1, Example 21.16] shows that for the absolute loss one cannot achieve such a result. But whether Theorem 2 holds for the hinge loss, which is a typical choice in classification, is unknown to us. We speculate that this is indeed true, as at least the related classification tasks, that use the 0–1 loss, cannot achieve a rate faster than $\frac{1}{\epsilon}$ [19, Theorem 6.8].

Corollary 2 sketches a scenario in which sample complexity improvements of MR can be at most a constant over their supervised counterparts. This may sound

like a negative result, as other methods with similar assumptions can achieve exponentially fast learning rates [16, Chapter 6]. But constant improvement can still have significant effects, if this constant can be arbitrarily large. If we set the regularization parameter μ in the concentric circles example high enough, the only possible classification functions will be the one that classifies each circle uniformly to one class. At the same time the pseudo-dimension of the supervised model can be arbitrarily high, and thus also the constant in Corollary 2. In conclusion, one should realize the significant influence constant factors in finite sample settings can have.

References

1. Anthony, M., Bartlett, P.L.: Neural Network Learning: Theoretical Foundations, 1st edn. Cambridge University Press, New York, USA (2009)
2. Azizyan, M., Singh, A., Wasserman, L.A.: Density-sensitive semisupervised inference. Computing Research Repository. abs/1204.1685 (2012)
3. Balcan, M., Blais, E., Blum, A., Yang, L.: Active property testing. In: 53rd Annual IEEE Symposium on Foundations of Computer Science, New Brunswick, NJ, USA, pp. 21–30 (2012)
4. Balcan, M.F., Blum, A.: A discriminative model for semi-supervised learning. J. ACM **57**(3), 19:1–19:46 (2010)
5. Bartlett, P.L., Bousquet, O., Mendelson, S.: Local Rademacher complexities. Ann. Stat. **33**(4), 1497–1537 (2005)
6. Belkin, M., Niyogi, P.: Towards a theoretical foundation for Laplacian-based manifold methods. J. Comput. Syst. Sci. **74**(8), 1289–1308 (2008)
7. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold regularization: a geometric framework for learning from labeled and unlabeled examples. JMLR **7**, 2399–2434 (2006)
8. Ben-David, S., Lu, T., Pál, D.: Does unlabeled data provably help? Worst-case analysis of the sample complexity of semi-supervised learning. In: Proceedings of the 21st Annual Conference on Learning Theory, Helsinki, Finland (2008)
9. Bholowalia, P., Kumar, A.: EBK-means: a clustering technique based on elbow method and k-means in WSN. Int. J. Comput. Appl. **105**(9), 17–24 (2014)
10. Boucheron, S., Bousquet, O., Lugosi, G.: Theory of classification: a survey of some recent advances. ESAIM Probab. Stat. **9**, 323–375 (2005)
11. Chapelle, O., Schölkopf, B., Zien, A.: Semi-Supervised Learning. The MIT Press, Cambridge (2006)
12. Darnstädt, M., Simon, H.U., Szörényi, B.: Unlabeled data does provably help. In: STACS, Kiel, Germany, vol. 20, pp. 185–196 (2013)
13. Globerson, A., Livni, R., Shalev-Shwartz, S.: Effective semisupervised learning on manifolds. In: COLT, Amsterdam, The Netherlands, pp. 978–1003 (2017)
14. Grandvalet, Y., Bengio, Y.: Semi-supervised learning by entropy minimization. In: NeuRIPS, Vancouver, BC, Canada, pp. 529–536 (2004)
15. Kloft, M., Brefeld, U., Laskov, P., Müller, K.R., Zien, A., Sonnenburg, S.: Efficient and accurate Lp-norm multiple kernel learning. In: NeuRIPS, Vancouver, BC, Canada, pp. 997–1005 (2009)
16. Mey, A., Loog, M.: Improvability through semi-supervised learning: a survey of theoretical results. Computing Research Repository. abs/1908.09574 (2019)
17. Mohri, M., Rostamizadeh, A., Talwalkar, A.: Foundations of Machine Learning. The MIT Press, Cambridge (2012)

18. Niyogi, P.: Manifold regularization and semi-supervised learning: some theoretical analyses. JMLR **14**(1), 1229–1250 (2013)
19. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, New York (2014)
20. Sindhwani, V., Niyogi, P., Belkin, M.: Beyond the point cloud: from transductive to semi-supervised learning. In: ICML, Bonn, Germany, pp. 824–831 (2005)
21. Sindhwani, V., Rosenberg, D.S.: An RKHS for multi-view learning and manifold co-regularization. In: ICML, Helsinki, Finland, pp. 976–983 (2008)
22. Vapnik, V.N.: Statistical Learning Theory. Wiley, Hoboken (1998)