Theoretical Computer Science ••• (••••) •••–•••

Contents lists available at ScienceDirect

Theoretical Computer Science

www.elsevier.com/locate/tcs

Localization of VC classes: Beyond local Rademacher complexities

N. Zhivotovskiy [a,b,*], S. Hanneke [c]

[a] *Skolkovo Institute of Science and Technology, Moscow, Russia*
[b] *Institute for Information Transmission Problems, Moscow, Russia*
[c] *Princeton, NJ 08542, USA*

ARTICLE INFO

Article history:
Received 15 January 2017
Received in revised form 22 August 2017
Accepted 7 November 2017
Available online xxxx

Keywords:
Statistical learning
PAC learning
Local metric entropy
Local Rademacher process
Shifted empirical process
Offset Rademacher process
ERM
Alexander's capacity
Disagreement coefficient
Massart's noise condition

ABSTRACT

In statistical learning the excess risk of empirical risk minimization (ERM) is controlled by $\left(\frac{\mathrm{COMP}_n(\mathcal{F})}{n}\right)^{\alpha}$, where n is a size of a learning sample, $\mathrm{COMP}_n(\mathcal{F})$ is a complexity term associated with a given class $\mathcal{F}$ and $\alpha \in [\frac{1}{2}, 1]$ interpolates between slow and fast learning rates. In this paper we introduce an alternative localization approach for binary classification that leads to a novel complexity measure: fixed points of the local empirical entropy. We show that this complexity measure gives a tight control over $\mathrm{COMP}_n(\mathcal{F})$ in the upper bounds under bounded noise. Our results are accompanied by a minimax lower bound that involves the same quantity. In particular, we practically answer the question of optimality of ERM under bounded noise for general VC classes.

1. Introduction

Since the early days of statistical learning theory understanding of the generalization abilities of empirical risk minimization has been a central question. In 1968, Vapnik and Chervonenkis [36] introduced the combinatorial property of classes of classifiers which we now call the *VC dimension*, which plays a crucial role not only in statistics but in many other areas of mathematics. By now it is strongly believed that the VC-dimension fully characterizes the properties of the empirical risk minimization algorithm. For example, when no restrictions are made on the distributions one can prove that the probability of error of the minimizer of empirical risk is close to the probability of error of the best classifier in the class, up to a term of order $\sqrt{\frac{d}{n}} + \sqrt{\frac{\log(\frac{1}{\delta})}{n}}$, with probability at least $1-\delta$, where d is the VC dimension of the class and n is the sample size. One can also prove a minimax lower bound (valid for any learning procedure) matching up to absolute constants. But the fact that VC dimension alone describes the complexity term appears to be true only in the agnostic case, when no assumptions are made on the labelling mechanism. It was noticed several times in the literature, that when considering bounded noise, VC dimension alone is not a right complexity measure of ERM [28,31,17]. Until now an exact right complexity measure has only been identified for a few specific classes. In this paper we propose a complexity measure which

* Corresponding author at: Skolkovo Institute of Science and Technology, Moscow, Russia.
E-mail addresses: nikita.zhivotovskiy@phystech.edu (N. Zhivotovskiy), steve.hanneke@gmail.com (S. Hanneke).

https://doi.org/10.1016/j.tcs.2017.12.029
0304-3975/

provides upper bounds on the risk of ERM, as well as lower bounds under regularity conditions, and therefore represents the right complexity measure for ERM in these cases.

In the last twenty years many efforts were made to understand the conditions that imply fast $\frac{1}{n}$ convergence rates, instead of slow $\frac{1}{\sqrt{n}}$ rates. By now these conditions are well understood; we refer for example to van Erven et al. [39] for an extensive survey and related results. At the beginning of the 2000s, so-called *localized* complexities (Bartlett et al. [5], Koltchinskii [21]) were introduced to statistical learning and became popular techniques for proving $\frac{1}{n}$ rates in different scenarios. But in addition to better rates, localization means that *only a small vicinity of the best classifier* really affects the learning complexity. Almost fifty years after the introduction of VC theory this phenomenon is still not fully understood and studied. Specifically, we lack tight error bounds based on localization and expressed in terms of intuitively-simple and calculable combinatorial properties of the class. Existing approaches based on localization (mainly, via *local Rademacher complexities*) are typically difficult to calculate directly, and the simpler relaxations of these bounds in the literature use localization merely to gain improvements due to the *noise conditions*, but fail to maintain the important improvements due to the local *structure of the function class* (i.e., localization of the complexity term in the bound). Moreover, to the best of our knowledge, in classification literature there are no known general minimax lower bounds in terms of localized processes.

There does exist one line of results which simultaneously give fast convergence rates and perform direct localization of a class of classifiers, to arrive at simple generalization bounds. Specifically, Massart and Nédélec [28] proved that under Massart's bounded noise condition, generalization of order $\frac{d}{nh}\log(\frac{nh^2}{d}) + \frac{\log(\frac{1}{\delta})}{nh}$ is possible, where h is a margin parameter responsible for the noise level. To derive this bound, Massart and Nédélec use a localized analysis to obtain improved rates under these noise conditions. However, the bound does not reflect this localization in the *complexity term* itself: in this case, the factor $d\log(\frac{nh^2}{d})$. Giné and Koltchinskii [13] refined this bound, establishing generalization of order $\frac{d}{nh}\log(\tau(\frac{d}{nh^2})) + \frac{\log(\frac{1}{\delta})}{nh}$ for empirical risk minimization, where τ is a distribution-dependent quantity they refer to as *Alexander's capacity function* (from the work of Alexander in the 80s [1]). Very recently, Hanneke and Yang [16] introduced a novel combinatorial parameter $\mathbf{s}$, called the *star number*, which gives perfectly-tight distribution-free control on $\tau(\frac{d}{nh^2})$, and generally cannot be upper bounded in terms of the VC dimension. Thus (as noted by Hanneke [17]), in terms of distribution-free guarantees on the generalization of empirical risk minimization, the implication of Giné and Koltchinskii's result is a bound $\frac{d}{nh}\log(\mathbf{s} \wedge \frac{nh^2}{d}) + \frac{\log(\frac{1}{\delta})}{nh}$. However, this bound is sometimes suboptimal. In this paper we will give a new argument showing potential gaps of this bound.

The aim of this paper is to perform a tight distribution-free localization for VC classes under bounded noise by introducing an appropriate distribution-free complexity measure, thus resolving the existing gap between upper and lower bounds. The complexity measure is a localized empirical entropy measure: essentially, a fixed point of the local empirical entropy. Most of the results will be proved in expectation and in deviation. Although results in expectation can usually be derived by integrating the results in deviation, we will directly prove results in expectation in the main part of the paper. Proofs of standard technical propositions and some results in deviation will be moved to the appendix. This paper is organized as follows:

- In section 2 we introduce the notation, definitions and previous results.
- In section 3 we introduce and further develop the machinery, based on the combination of shifted empirical processes [24] and offset Rademacher complexities [26]. We also obtain a new upper bound on the error rate of empirical risk minimization in the realizable case, involving the star number and the growth function, which refines a recent result of Hanneke [17] in some cases; this bound is a strict improvement over the distribution-free bound implied by the result of Giné and Koltchinskii in the realizable case.
- Section 4 is devoted to an upper bound in terms of fixed point of global metric entropy. Although it gives a fast convergence rate $\frac{1}{n}$, it involves only a global information about the class. Thus, this bound is suboptimal in some interesting cases, as are the other bounds in the literature based solely on global complexities for the class. We include the proof nevertheless, as it cleanly illustrates certain aspects of our approach; for simplicity, we only present this result in the realizable case.
- Section 5 contains our main results. In this section we introduce the local empirical entropy and prove that fixed points of local empirical entropy control the complexity of ERM under bounded noise.
- Section 6 is devoted to a novel lower bound in terms of fixed points of local empirical entropy under mild regularity assumptions.
- Section 7 contains examples of values of fixed points for some standard classes.
- Section 8 is devoted to discussions and some related general results. Specifically, we prove that bounds based on our complexity measure are always not worse than the bounds based on local Rademacher (Gaussian) complexities.

2. Notation and previous results

We define the *instance space* $\mathcal{X}$ and the *label space* $\mathcal{Y} = \{1, -1\}$. We assume that the set $\mathcal{X} \times \mathcal{Y}$ is equipped with some σ-algebra and a probability measure P on measurable subsets is defined. We also assume that we are given a set of classifiers $\mathcal{F}$; these are measurable functions with respect to the introduced σ-algebra, mapping $\mathcal{X}$ to $\mathcal{Y}$. We may always

Please cite this article in press as: N. Zhivotovskiy, S. Hanneke, Localization of VC classes: Beyond local Rademacher complexities, Theoret. Comput. Sci. (2018), https://doi.org/10.1016/j.tcs.2017.12.029