GML Project Proposal

Zhihan Jin
pascalprimer@gmail.com

Matteo Russo
mrusso@eth.ch

March 17, 2021

1 Paper Summary

We are reviewing the paper [BHMZ20] by Bousquet et al. and trying to find new directions of possible research. Recently, in a result by Hanneke et al. in [Han16a], a bound on the optimal PAC learning sample complexity of concept class $\mathbb{C}$ has been found:

$$\Theta\left(\frac{1}{\varepsilon}\left(\mathrm{d}_{\mathrm{VC}}+\log\frac{1}{\delta}\right)\right)$$

The caveat is that the estimator achieving such a bound belongs to the family of improper learning estimators (classifiers/algorithms), which in words means that it may output an hypothesis not belonging to the concept class. Their main contribution in this paper can be outlined in two main macro-points:

1. Establishing the conditions for which a proper learner achieves the same bound, which, in fact, boils down to bounding the dual Helly number.
2. Establishing the optimality of SVM learners when the concept class is that of linear separators, a long sought conjecture posed originally by Vapnik and Chervonenkis in [VC74].

1.1 Proper vs Improper learning

We now proceed in defining proper and improper learning: consider an estimated (learnt) classifier $\hat{h}_n : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{X}$ is a measurable space representing the support of distribution (probability measure) $\mathcal{P}$ of the the n samples $x_1, \ldots, x_n$ we draw, and $\mathcal{Y} \triangleq \{-1, 1\}$.

- Proper learning: the estimated (learnt) classifier is said to be proper if $\hat{h}_n \in \mathbb{C}$.
- Improper learning: the estimated (learnt) classifier is said to be improper if $\hat{h}_n \notin \mathbb{C}$.

1.2 Empirical Risk Minimizer (ERM) and beyond

In Fig. 1, we list several related results. An important direction is to exclude the $\log\frac{1}{\varepsilon}$ in the classical bound for ERM, i.e. approaching the optimal sampling complexity for VC classes. However, there is still a large constant, relating to the VC class, left in the bounds by Bousquet et al. [BHMZ20].

Bounds on the sample complexity of PAC learning		
Improper Learning	$\Theta\left(\frac{\mathrm{d_{VC}}}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	[Han16a] [EHKV89]
Any ERM	$O\left(\frac{\mathrm{d_{VC}}}{\varepsilon}\log(\frac{1}{\varepsilon}\wedge\frac{\mathfrak{s}}{\mathrm{d_{VC}}})+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{\mathrm{d_{VC}}}{\varepsilon}+\frac{1}{\varepsilon}\log(\frac{1}{\varepsilon}\wedge\mathfrak{s})+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	[Han16b] [VC74]
Proper Learning $\mathcal{M}_{\mathrm{prop}}(\varepsilon,\delta)$	$O\left(\frac{\mathrm{d_{VC}}\mathrm{k}_w^2}{\varepsilon}\log(\mathrm{k}_w)+\frac{\mathrm{k}_w^2}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{\mathrm{d_{VC}}}{\varepsilon}+\frac{1}{\varepsilon}\log(\mathrm{k}_w)+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	[BHMZ20]
SVM / Halfspaces in $\mathbb{R}^p$	$\Theta\left(\frac{n}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	[BHMZ20]
Maximum Class (Proper)	$\Theta\left(\frac{\mathrm{d_{VC}}}{\varepsilon}+\frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	[BHMZ20]

Figure 1: Summary of results on the sample complexity of (ϵ,δ)-PAC learning. $\mathrm{d_{VC}}$ denotes the *VC dimension* , $\mathfrak{s}$ the *star number* , and k_w the *dual Helly number* discussed in this article. Specific definitions, conditions, and ranges of parameters for which the results hold are discussed below.

1.3 Dual Helly number and its variants

For any configuration $S\subset\mathcal{X}\times\mathcal{Y}$, define $\mathbb{C}[S]\triangleq\{h\in\mathbb{C}:\forall(x,y)\in S,h(x)=y\}$, the set of feasible concepts. For a finite multiset $\mathbb{C}'\subseteq\mathbb{C}$ let $\mathrm{Majority}(\mathbb{C}'):\mathcal{X}\to\{0,1,?\}$ denote the majority-vote classifier defined by:

$$\mathrm{Majority}(\mathbb{C}')(\mathrm{x})=\begin{cases}0 & \left|\{c\in\mathbb{C}':c(x)=0\}\right|>\frac{|\mathbb{C}'|}{2},\\ 1 & \left|\{c\in\mathbb{C}':c(x)=1\}\right|>\frac{|\mathbb{C}'|}{2},\\ ? & \text{otherwise.}\end{cases}$$

For $\ell\geq 2$, define the set $\mathcal{X}_{\mathbb{C}',\ell}\subseteq\mathcal{X}$ of all the points x on which less than $\frac{1}{\ell}$-fraction of all classifiers in $\mathbb{C}'$ disagree with the majority. That is,

$$\mathcal{X}_{\mathbb{C}',\ell}=\left\{x\in\mathcal{X}:\sum_{h\in\mathbb{C}'}\mathbb{1}_{h(x)\neq h_{\mathrm{maj}}(x)}<\frac{|\mathbb{C}'|}{\ell}\right\},$$

where $h_{\mathrm{maj}}=\mathrm{Majority}(\mathbb{C}')$.

- Dual Helly number k_w of $\mathbb{C}$ is the smallest $k\in\overline{\mathbb{N}}$ such that whenever $\mathbb{C}[S]=\emptyset$, there exists some $W\subset_{\leq k}S$ with $\mathbb{C}[W]=\emptyset$. Briefly, k_w indicates the minimum certificate of each unrealizable configure $S\subset\mathcal{X}\times\mathcal{Y}$.

- Star Number $\mathfrak{s}$ of $\mathbb{C}$ is the maximum size of a realizable configuration $S\subset\mathcal{X}\times\mathcal{Y}$ such that every neighbour (from flipping a single y) of S is also realizable. Note that $\mathfrak{s}$ is often ∞.

- Hollow Star Number k_o of $\mathbb{C}$ is the maximum size of an unrealizable configuration $S\subset\mathcal{X}\times\mathcal{Y}$ such that every neighbour (from flipping a single y) of S is realizable. Trivially, $\mathfrak{s}\geq\mathrm{k}_o-1$.

- Projection Number k_p of $\mathbb{C}$ is the smallest $k\in\overline{\mathbb{N}}$ such that for every finite multiset $\mathbb{C}'\subset\mathbb{C}$, there exists some $h\in\mathbb{C}$ agreeing with $\mathrm{Majority}(\mathbb{C}')$ on $\mathcal{X}_{\mathbb{C}',k}$. This indicates the realizability of highly agreed configuration (data) points.

Lemma 1.1. *The two following results about star numbers hold:*

- $\mathrm{k}_o \leq \mathrm{k}_p \leq \mathrm{k}_w$*;*
- $\mathrm{k}_o = \mathrm{k}_p = \mathrm{k}_w$ *when* $\mathrm{k}_w < \infty$ *or* $\mathbb{C}$ *is closed.*

1.4 Result I: Upper bound on proper learning classes

The goal here is to understand whether and when there exists a proper learning classifier/algorithm achieving the same optimal PAC learning sample complexity as the improper learner shown in [Han16a]. The authors show a relatively simple algorithm that based on majority voting and projection back to the concept class achieves the following sample complexity order:

$$\mathcal{M}_{\text{prop}}(\varepsilon, \delta) \in O\left(\frac{\mathrm{k}_p^3}{\varepsilon}\left(\mathrm{d}_{\text{VC}} \log \mathrm{k}_p + \log \frac{1}{\delta}\right)\right)$$

We note that for bounded projection number k_p, the sample complexity is of the same order as for the improper learner. For a clear and compact summary, please refer to the third row in Fig. 1. The proof of this fact derives from a slight modification of the "conditioning" argument used in [Han16a] to prove optimality (in PAC learning terms) of the improper learner.

1.5 Result II: Lower bound on proper learning classes

The goal here is to understand what is the best sample complexity any proper learning classifier/algorithm with bounded and unbounded hollow star number is. Reminiscent of the useful upper bound of above, Bousquet et al. prove the following interesting two results on the sample complexity of a proper learner.

$$\mathcal{M}_{\text{prop}}(\varepsilon, \delta) \in \Omega\left(\frac{1}{\varepsilon}\left(\mathrm{d}_{\text{VC}} + \log \mathrm{k}_o + \log \frac{1}{\delta}\right)\right) \qquad \text{if } \mathrm{k}_o < \infty \text{ and } \varepsilon \leq \frac{1}{\mathrm{k}_o}$$

$$\mathcal{M}_{\text{prop}}(\varepsilon, \delta) \notin o\left(\frac{1}{\varepsilon} \log \frac{1}{\delta}\right) \qquad \text{if } \mathrm{k}_o = \infty$$

For a clear and compact summary, please refer to the third row in Fig. 1. The proof of these facts comes from a coupon collector argument and we refer to [BHMZ20] for complete details.

1.6 Result III: Stable compression schemes

We can define a *compression scheme* as a pair of functions (κ, ρ) called compression and reconstruction functions respectively such that for realizable $S \in \mathcal{X} \times \mathcal{Y}$, we get that there exists a subset $\kappa(S) \subseteq S$, with $|\kappa(S)| \leq \ell$, for which $\rho(\kappa(S))$ is a classifier correct on S. The general sample complexity of a compression scheme can be shown to be

$$\mathcal{M}_{(\kappa,\rho)}(\varepsilon, \delta) \in \Theta\left(\frac{1}{\varepsilon}\left(\ell \log \frac{1}{\varepsilon} + \log \frac{1}{\delta}\right)\right)$$

We can note that this is not necessarily better than the general ERM bound as, for instance, in the the case of SVM, for which $\ell = p + 1$. Nevertheless, for SVM in particular, we know that removing any non support vector from the sample points in space does not change the classifier: this motivates the following definition.

A *stable compression scheme* is a compression scheme such that the choice of the compression set does not change if the whole data sample is considered or if a sub-sample consisting only of the compression points is considered. Formally, for any $(x, y) \in S \setminus \kappa(S)$,

$$\kappa(S) = \kappa(S \setminus \{(x, y)\})$$

Then, one can establish that the sample complexity of a (κ, ρ) stable compression scheme of size ℓ is

$$\mathcal{M}_{(\kappa,\rho)}(\varepsilon, \delta) \in \Theta\left(\frac{1}{\varepsilon}\left(\ell + \log\frac{1}{\delta}\right)\right)$$

Hereby, we note that the $\log\frac{1}{\varepsilon}$ factor with respect to a general compression scheme has vanished.

2 Possible directions

2.1 Open questions from the paper

The paper poses two important and interesting questions that seem very general:

1. Is the sample complexity always characterized by the hollow star number k_o and the VC dimension d_{VC}?
2. Is the optimal size of a proper stable compression scheme always characterized by the hollow star number k_o and the VC dimension d_{VC}?

2.2 Connection between Helly-similar numbers and VC dimension

This direction is fundamentally "orthogonal" to the others we consider here. In particular, we would like to be able to relate the VC dimension d_{VC} with the dual Helly number k_w (or its variants) in order to be able to express sample complexity as a function of only one parameter. This quest may be very challenging in general but could show, if true, a very interesting and fascinating result per se: a quantity related to the measure of sets intersection in combinatorial geometry and a measure of estimators complexity are one a function of the other.

2.3 Bridging the gap between upper and lower bound

The upper and lower bounds suggested by Bousquet et al. [BHMZ20] as listed in Fig. 1 successfully exclude the log factor but there is still a $\mathsf{poly}(\mathrm{k}_w)$ gap between them. We hope to understand the inner reason in two aspects:

- Are there cases matching both bounds?
- What is the barrier in the proof that gives this gap?

It would be fortunate and exciting if we could improve either bounds.

Also, the lower bounds confirms the necessity of a log factor when $\mathrm{k}_o = \infty$. Another important thing is to understand when this happens or at least characterize several commonly used concept classes with $\mathrm{k}_o = \infty$.

2.4 Specializing to a new Stable Compression Scheme

For this particular direction, we would like to find a new stable compression scheme and establish optimal sample complexity PAC learning bounds. For a concrete example, we can consider a parallel to the SVM in the following sense: the SVM has been to be optimal for linear separators concept class. Hereby, the question would be whether there exists an equivalent-to-SVM stable compression learner such that, for the concept class of polynomial separators, it achieves optimal sample complexity and this depends on the VC dimension d_{VC} as well as on the hollow star number k_o. In other words, it would be a specialization of question 2 from the suggested open questions 2.1 (and perhaps more accessible).

References

[BHMZ20] Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound. In Jacob Abernethy and Shivani Agarwal, editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 582–609. PMLR, 09–12 Jul 2020. 1, 1.2, 1.5, 2.3

[EHKV89] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989. 1.2

[Han16a] Steve Hanneke. The optimal sample complexity of pac learning. *Journal of Machine Learning Research*, 17(38):1–15, 2016. 1, 1.2, 1.4

[Han16b] Steve Hanneke. Refined error bounds for several learning algorithms. *Journal of Machine Learning Research*, 17(135):1–55, 2016. 1.2

[VC74] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian]*. Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979). 2, 1.2