Information bounds for learners with bounded VC dimension

Gregoire Fournier

March 3, 2025

1 Notations

Set $Z = \mathcal{X} \times \mathcal{Y}$ the examples domain and $\mathcal{H} = \{h_w : w \in \mathcal{W}\}$ the hypothesis set indexed by $\mathcal{W}$. A loss function $\ell : \mathcal{H} \times Z \to \mathbb{R}^+$. The training set $S = S_n = (Z_1, ..., Z_n)$ is constituted of n iid samples from Z with distribution μ. A proper learning algorithm $\mathcal{A}$ picks $h_W \in \mathcal{H}$ according to $P_{W|S}$. For any $w \in \mathcal{W}$ define:

$$L_\mu(w) = \mathbb{E}[\ell(h_w, Z)],\ Z \sim \mu$$

$$L_S(w) = \frac{1}{n}\sum_i \ell(h_w, Z_i),\ Z_i \sim \mu$$

$$gen_\mu(w) = L_\mu(w) - L_S(w)$$

$$gen_\mu(\mathcal{A}) = gen_\mu(P_{W|S}) = \mathbb{E}[L_\mu(w) - L_S(w)]$$

In the last definition the expectation is taken with regards to $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$. Let the super-sample $\mathcal{Z}$ be an $2\times n$ array of i.i.d random variables following μ, U be a sequence of i.i.d. Bernoulli random variables in $\{0, 1\}$, independent from $\mathcal{Z}$, with $\mathbb{P}(U_i = 0) = \mathbb{P}(U_i = 1) = 1/2$, and, for every $n \in \mathbb{N}$, let $S_n = (\mathcal{Z}_{Uj,j})_{1\le j\le n}$. For an algorithm $\mathcal{A}$, define the conditional mutual information of $\mathcal{A}$, denoted $\mathrm{CMI}_\mu(\mathcal{A})$, to be the conditional mutual information $I(\mathcal{A}(S); U|\mathcal{Z})$. Define the distribution-free CMI as $\mathrm{CMI}(\mathcal{A}) = sup_{\mathcal{Z}} I(\mathcal{A}(S); S)$.

2 Presentation of the problem

We are interested in the following conjectures[1]:

Conjecture 1. There is a constant $c > 0$ such that, for every VC class $\mathcal{H}$, with dimension d, if there exists a proper learning algorithm with the expected risk no greater than cd/n for every realizable distribution μ, then there exists a proper learning algorithm $\mathcal{A}_n$ with $\mathrm{CMI}_\mu(\mathcal{A}_n) \le cd$ and $\mathbb{E}(L_{S_n}(\mathcal{A}_n(S_n)) \le cd/n$ for every realizable distribution μ.

Conjecture 2.There is a constant $c > 0$ such that, for every VC class $\mathcal{H}$, with dimension d, there exists a (possibly improper) learning algorithm $\mathcal{A}_n$ such that $\mathrm{CMI}_D(\mathcal{A}_n) \le cd$ and $\mathbb{E}(L_{S_n}(\mathcal{A}_n(S_n)) \le cd/n$ for every realizable distribution μ.

3 Some results for bounded VC-dimension

In this part $\mathcal{Y}$ is restricted to $\{-1, 1\}$. First we justify why bounded VC-dimension yields good generalization properties [2]:

Definition 1 (Empirical Rademacher complexity). Let σ be a vector of independent uniform random variables taking values in $\{-1, 1\}$. The empirical Rademacher complexity of $\mathcal{H}$ with respect to the sample S is defined as:

$$\hat{\mathcal{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma[sup_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i h(Z_i)]$$

Theorem 2 (Rademacher complexity bounds – binary classification). Let $\mathcal{H}$ be a family of functions taking values in $\{-1, 1\}$ and let μ be the distribution over the input space. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over a sample $S = S_n = (Z_1, ..., Z_n)$ drawn according to μ, then for all $w \in \mathcal{W}$:

$$gen_\mu(w) \leq \hat{\mathcal{R}}_S(\mathcal{H}) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

When the VC-dimension if finite, we can bound the Rademacher complexity to the following corollary:

Corollary 3. Let $\mathcal{H}$ be a family of functions taking values in $\{-1, 1\}$ with VC-dimension d. Then, for any $\delta > 0$, with probability at least $1 - \delta$, for all $w \in \mathcal{W}$:

$$gen_\mu(w) \leq \sqrt{\frac{2d\ \log(\frac{en}{d})}{n}} + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

Recently progress has been made regarding the bounding of generalization using CMI, as avoiding over-fitting intuitively seem linked to low CMI.[3]

Theorem 4 (Generalization from CMI). Let $Z = \mathcal{X} \times \{0, 1\}$, for an algorithm $\mathcal{A}$ and $S = S_n = (Z_1, ..., Z_n)$:

$$gen_\mu(\mathcal{A}) \leq \sqrt{\frac{2\ \mathrm{CMI}_\mu(\mathcal{A})}{n}}$$

Theorem 5 (Bounding CMI). Let $Z = \mathcal{X} \times \{0, 1\}$ and $\mathcal{H}$ a hypothesis class with VC dimension d. Then, there exists an empirical risk minimizer $\mathcal{A}$ such that $\mathrm{CMI}(\mathcal{A}) \leq d\ \log n + 2$.

Note that we only presented upper bounds of the generalization error. It is also possible to derive lower bounds, using probabilistic method types of arguments.

4 On the optimal complexity of PAC learning

In this part $\mathcal{Y}$ is restricted to $\{-1, 1\}$. To better understand the question of information for good generalisation algorithms, it is useful to mention the optimal complexity of PAC learning depending on the nature of the algorithm and on the type of problem [4]:

Bounds on the sample complexity of PAC learning		
Improper Learning	$\Theta\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Hanneke2016OP[5] Ehrenfeucht89[6]
Any ERM	$O\left(\frac{d}{\varepsilon}\log(\frac{1}{\varepsilon} \wedge \frac{s}{d}) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log(\frac{1}{\varepsilon} \wedge \frac{s}{d}) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Hanneke2016RE[7] Vapnik74[8]
Proper Learning	$O\left(\frac{dk^2}{\varepsilon}\log(k) + \frac{k^2}{\varepsilon}\log\frac{1}{\delta}\right)$ $\Omega\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log(k) + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Bousquet20[4]
SVM in $\mathbb{R}^n$	$\Theta\left(\frac{n}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Bousquet20[4]
Maximum Class	$\Theta\left(\frac{d}{\varepsilon} + \frac{1}{\varepsilon}\log\frac{1}{\delta}\right)$	Bousquet20[4]

Table 1: Summary of results on the sample complexity of (ε, δ)-PAC learning. d denotes the VC dimension, s the star number [9], and k the dual Helly number.

We choose not to develop every bound in the table, but instead focus on the upper bounds for proper and improper learning. They both rely on a majority classifier. The majority is composed of ERMs trained on different overlapping subsets. In the case of obtaining the mentioned improper learning upper bounds, the subsets are determined as follows [5]:

Algorithm $\mathcal{A}(S; T)$

0. If $|S| \leq 3$
1. Return $\{S \cup T\}$
2. Let S_0 denote the first $|S| - 3|S|/4$ elements of S, S_1 the next $|S|/4c$ elements, S_2 the next $|S|/4c$ elements, and S_3 the remaining $|S|/4c$ elements
3. Return $\mathcal{A}(S_0; S_2 \cup S_3 \cup T) \cup \mathcal{A}(S_0; S_1 \cup S_3 \cup T) \cup \mathcal{A}(S_0; S_1 \cup S_2 \cup T)$

The main idea is given two classifiers, each consistent with an i.i.d. data set independent from the other. The probability that they both make a mistake on a random point is bounded by bounding the error rate of the first classifier, then bounding the error rate of under the conditional distribution given that the first one made a mistake.

Note that the final algorithm takes the majority of the classifiers, hence giving an improper algorithm. Following this approach, projecting a majority classifier into the proper space yielded the upper bound in Table 1.

5 Proper Learning, Helly Number, and an Optimal SVM Bound

In this part $\mathcal{Y}$ is restricted to $\{-1, 1\}$. There have been several reductions made from the PAC learning main problem to help understand the optimal sample complexity of some classes. In [4] are introduced geometrical measurement of the complexity of the space, namely: the dual Helly number, the hollow star number and the projection number. We choose to focus on the projection number as it is the one that better suits our application of the study of improper versus proper learning.

For any finite $\mathcal{H}' \subseteq \mathcal{H}$ and $l \geq 2$, define the set $\mathcal{X}_{\mathcal{H}',l} \subseteq \mathcal{X}$ of all the points x on which less than $1 - l$ fraction of all classifiers in $\mathcal{H}'$ disagree with the majority classifier h_{maj}:

$$\mathcal{X}_{\mathcal{H}',l} = \{x \in \mathcal{X} : \sum_{h \in \mathcal{H}'} \mathbb{1}[h(x) \neq h_{maj}(x) < \frac{|\mathcal{H}'|}{l}]\}$$

We denote by $Proj_{\mathcal{H}}(\mathcal{H}')$ any element in $\{h \in \mathcal{H} : h(x) = \text{Majority}(\mathcal{H}'), \text{ for all } x \in \mathcal{X}_{\mathcal{H}',k)p}\}$.

Definition 6 (Projection Number). The projection number of a class $\mathcal{H}$, denoted by k_p, is the smallest integer $k \geq 2$ such that, for any finite $\mathcal{H}' \subseteq \mathcal{H}$, there exists $h \in \mathcal{H}$ that agrees with Majority($\mathcal{H}'$) on the entire set $\mathcal{X}_{\mathcal{H}',k}$. If no such integer k exists, define $k_p := \infty$.

The algorithm for generating the subsets is as follows:

Algorithm $\mathcal{A}(S; T)$

1. If $|S| \leq 4$, Return ERM($S \cup T$)
2. Let S_0 denote the first $|S|/2$ elements of S
3. Let $S_1, ...S_{k_p+1}$ be independent uniform without replacement subsamples of $S \setminus S_0$ of size $|S|/4$
3. Let $h_i = \mathcal{A}(S_0; T \cup S_i)$ for each $i = 1, ..., k_p + 1$
4. Return $\hat{h} = Proj_{\mathcal{H}}(h_1, \ldots, h_{k_p+1})$

To better understand the complexity indicators introduced, we will introduce one more lemma. We denote by $\mathcal{H}[S]$ the subset of $\mathcal{H}$ which elements are consistent with S.

Definition 7 (The dual Helly number). The dual Helly number of $\mathcal{H}$, denoted by k_w, as the smallest integer k such that, for any S sampled from Z such that $\mathcal{H}S = \emptyset$, there is a set $W \subseteq S$ with $|W| \leq k$ such that $\mathcal{H}[W] = \emptyset$. If no such k exists, we define $k_w = \infty$.

Definition 8 (The hollow star number). The hollow star number of $\mathcal{H}$, denoted by k_o, as the largest integer k such that there is a sequence $S = ((x_1, y_1), \ldots (x_k, y_k)) \in (\mathcal{X} \times \mathcal{Y})^k$ for which $\mathcal{H}[S] = \emptyset$ and that every neighbor S' of S verifies $\mathcal{H}[S'] \neq \emptyset$. If no such largest k exists, define $k_o = \emptyset$.

Lemma 9.

- $k_o \leq k_p \leq k_w$.
- If k_w is finite or $\mathcal{H}$ closed, $k_o = k_p = k_w$.

6 Bounding generalization error via Mutual information

Now we turn our attention to input-output mutual information of algorithms. We present various generalization guarantees for learning algorithms that are stable in mutual information [10][11].

Definition 10. A random process $\{X_t\}_{t\in T}$ along with a metric d on T is subgaussian if $\mathbb{E}X_t = 0$ for all t and:

$$\mathbb{E}[e^{\lambda(X_t - X_s)}] \leq e^{\frac{1}{2}\lambda^2 d^2(t,s)}, \text{ for all } \lambda \geq 0, t, d \in T$$

Using Hoeffding inequality, we get that $\{gen_\mu(w)\}_{w\in\mathcal{W}}$ is a gaussian sub-processes with the metric $d(gen_\mu(w), gen_\mu(v)) = \frac{||\ell(h_w,\cdot)-\ell(h_v,\cdot)||_\infty}{\sqrt{n}}$ for any distribution μ on Z.

Consider a pair of random variables X and Y with joint distribution $P_{X,Y}$. Let $\bar{X}, \bar{Y}$ be independent copies of X, Y such that $P_{\bar{X},\bar{Y}} = P_X \otimes P_Y$.

Lemma 11. If $f(\bar{X}, \bar{Y})$ is σ-subgaussian under $P_{\bar{X},\bar{Y}} = P_X \otimes P_Y$,
i.e for all λ, $\mathbb{E}(e^{\lambda(f(\bar{X},\bar{Y})-\mathbb{E}(f(\bar{X},\bar{Y})}) \leq e^{\frac{1}{2}\lambda^2\sigma^2}$, then:

$$|\mathbb{E}[f(X,Y)] - \mathbb{E}[f(\bar{X},\bar{Y})| \leq \sqrt{2\sigma^2 I(X;Y)}$$

Observe that $gen_\mu(P_{W|S})$ can be written as $\mathbb{E}[f(\bar{S},\bar{W}) - \mathbb{E}[f(S,W)]$ where the joint distribution of S and W is $P_{S,W} = \mu^{\otimes n} \otimes P_{W|S}$ and $f(s,w) = \frac{1}{n}\sum_i \ell(w, z_i)$.
If $\ell(w, Z)$ is σ-subgaussian for all $w \in W$, then $f(S, w)$ is $\sigma/\sqrt{n}$-subgaussian due to the i.i.d. assumption on Z_i's, hence $f(\bar{S},\bar{W})$ is $\sigma/\sqrt{n}$-subgaussian.
Using the previous lemma, we get upper bounds on the generalisation error:

Theorem 12. Suppose $\ell(w, Z)$ is a σ-subgaussian under μ for all $w \in W$ then:

$$|gen_\mu(P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n} I(S;W)}$$

Define the collection of empirical risks of the hypotheses in W, $\Lambda_W(S) = (L_S(w))w \in W$. Setting X to $\Lambda_W(S)$, Y to W and picking $f(\Lambda_W(s), w) = L_s(w)$ in the previous lemma yields:

Theorem 13. Suppose $\ell(w, Z)$ is a σ-subgaussian under μ for all $w \in W$ then:

$$|gen_\mu(P_{W|S})| \leq \sqrt{\frac{2\sigma^2}{n} I(\Lambda_W(S);W)}$$

Note that it is a slight improvement as $I(\Lambda_W(S);W) \leq I(S;W)$ as $\Delta_W(S) - S - W$ is a Markov chain, seeing that for all $w \in W$, $L_S(w)$ is a function of S.

This analysis with a bit more work adds a concentration bound for the absolute generalization error, $gen^+_\mu(\mathcal{A}) = gen^+_\mu(P_{W|S}) = \mathbb{E}[|L_\mu(w) - L_S(w)|]$.

Theorem 14. Suppose $\ell(w, Z)$ is a σ-subgaussian under μ for all $w \in W$. If a learning algorithm satisfies that $I(\Lambda_W(S);W) \leq \varepsilon$, then:

$$gen^+_\mu(P_{W|S}) \leq \sqrt{\frac{2\sigma^2}{n}(\varepsilon + \log 2)}$$

7 Chaining method for Mutual information

Another way to get bounds on the generalization is to exploit the dependencies between hypotheses. The technique of chaining is known to give tighter bounds than the union bound. We start with a fundamental result which is based on the chaining method [12][13]. A random process $\{X_t\}_{t\in T}$ is separable if there is $T_0 \subseteq T$ countable such that $X_t \in \lim_{\substack{s\to t \\ s\in T_0}} X_s$ almost surely, i.e $\exists (s_n)_n$ sequence in T_0 such that $s_n \to t$ and $X_{s_n} \to X_t$. For a metric space (T,d), let $N(T,d,\varepsilon)$ denote the covering number of (T,d) at scale ε, similar to an ε-net.

Theorem 15. Assume that $\{X_t\}_{t\in T}$ is a separable subgaussian process on the bounded metric space (T,d), then:

$$\mathbb{E}[sup_{t\in T} X_t] \leq 6 \sum_{k\in\mathbb{Z}} 2^{-k} \sqrt{\log\ N(T,d,2^{-k})}$$

An ϵ-partition $\mathcal{P} = \{A_1, A_2, ..., A_m\}$ of the set T of the metric space (T,d) verifies that for all $i \in [m]$, A_i can be contained within a ball of radius ϵ.

Definition 16. A sequence of partitions $\{\mathcal{P}_k\}_{k=m}^{\infty}$ of a set T is called an increasing sequence if for all $k \geq m$ and each $A \in \mathcal{P}_{k+1}$, there exists $B \in \mathcal{P}_k$ such that $A \subseteq B$.

For any such sequence and any $t \in T$, let $[t]_k$ denote the unique set $A \in \mathcal{P}_k$ such that $t \in A$. Now we introduce the chaining method for mutual information [13]:

Theorem 17. Assume that $\{X_t\}_{t\in T}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W},d)$. Let $\{\mathcal{P}_k\}_{k=k_1(T}^{\infty}$ be an increasing sequence of partitions of T, where for each $k \geq k_1(T)$, $\mathcal{P}_k$ is a 2^{-k}-partition of (T,d).

$$\mathbb{E}[X_W] \leq 3\sqrt{2} \sum_{i=k_1(T)}^{\infty} 2^{-k} \sqrt{I([W]_k; X_T)}$$

Furthermore, assume that $\{gen_\mu(w)\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W},d)$. Let $\{\mathcal{P}_k\}_{k=k_1(\mathcal{W})}^{\infty}$ be an increasing sequence of partitions of $\mathcal{W}$, where for each $k \geq k_1(\mathcal{W})$, $\mathcal{P}_k$ is a 2^{-k}-partition of $(\mathcal{W},d)$.

$$gen_\mu(P_{W|S}) \leq 3\sqrt{2} \sum_{i=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I([W]_k; S)}$$

If $0 \in \{\ell(h_w, .) : w \in W\}$, we can derive an upper bound for the absolute generalization error:

$$gen_\mu^+(P_{W|S}) \leq 3\sqrt{2} \sum_{i=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I([W]_k; S) + \log\ 2}$$

8 Chaining method for Conditional Mutual information

We can apply the chaining procedure to mutual information as the work in [14].

Theorem 18. Assume that $X_{\mathcal{W}} = \{X_w\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W}; d)$. Consider the sequence of functions $(\Pi_k)_{k\geq k_1(\mathcal{W})}$ where $k_1(\mathcal{W})$ is the largest integer that satisfies $2^{-(k_1-1)} \geq \text{diam}(\mathcal{W})$, and for all $k > k_1$, $\Pi_k : \mathcal{W} \to \mathcal{W}$ is a function satisfying $d(w; \Pi_k(w)) \leq 2^{-k}$.
Define $\tilde{W}_k = \Pi_k(W)$ for $k \geq k_1$ and $\tilde{W}_{k_1-1} = w_0$ for an arbitrary $w_0 \in \mathcal{W}$. We have:

$$\mathbb{E}[X_W] \leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\tilde{W}_{k-1}, \tilde{W}_k; X_{\mathcal{W}})}$$

Considering a X_W of interest and after removing the dependence in $Wk-1$, they present the following theorem.

Theorem 19. Assume that $X_{\mathcal{W}} = \{\sqrt{n}gen(w)\}_{w\in\mathcal{W}}$ is a separable subgaussian process on the bounded metric space $(\mathcal{W}; d)$ and the learned hypothesis W is a deterministic function of $X_{\mathcal{W}}$. Consider the sequence of functions $(\Pi_k)_{k\geq k_1(\mathcal{W})}$ where $k_1(\mathcal{W})$ is the largest integer that satisfies $2^{-(k_1-1)} \geq \text{diam}(\mathcal{W})$, and for all $k > k_1$, $\Pi_k : \mathcal{W} \to \mathcal{W}$ is a function satisfying $d(w; \Pi_k(w)) \leq 2^{-k}$.
Define $\tilde{W}_k = \Pi_k(W)$ for $k \geq k_1$. We have:

$$\mathbb{E}[X_W] \leq \frac{1}{\sqrt{n}} 6\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\tilde{W}_k; X_{\mathcal{W}})}$$

The proof starts as follows:

$$\begin{aligned}
\mathbb{E}[X_W] &\leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\tilde{W}_{k-1}, \tilde{W}_k; X_{\mathcal{W}})} \\
&= 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\tilde{W}_k; X_{\mathcal{W}}) + I(\tilde{W}_{k-1}; X_{\mathcal{W}}|\tilde{W}_k)} \\
&\leq 3\sqrt{2} \sum_{k=k_1(\mathcal{W})}^{\infty} 2^{-k} \sqrt{I(\tilde{W}_k; X_{\mathcal{W}}) + I(\tilde{W}_{k-1}; X_{\mathcal{W}})} \qquad (1)
\end{aligned}$$

(1) comes from the fact that if X, Y, Z form a Markov chain (in any order), then $I(X; Y|Z) \leq I(X; Y)$.
$\tilde{W}_{k-1} \perp \tilde{W}_k | X_{\mathcal{W}}$, because of the deterministic relation between W and $X_{\mathcal{W}}$.

9 Reducing the CMI of an algorithm

In this part $\mathcal{Y}$ is restricted to $\{-1, 1\}$and we consider the $0-1$ loss.

Conjecture 1. There is a constant $c > 0$ such that, for every VC class $\mathcal{H}$, with dimension d, if there exists a proper learning algorithm with the expected risk no greater than cd/n for every realizable distribution μ, then there exists a proper learning algorithm $\mathcal{A}_n$ with $\mathrm{CMI}_\mu(\mathcal{A}_n) \leq cd$ and $\mathbb{E}(L_{S_n}(\mathcal{A}_n(S_n)) \leq cd/n$ for every realizable distribution μ.

Let the super-sample $\mathcal{Z}$ be an $2 \times n$ array of i.i.d random variables following μ, U be a sequence of i.i.d. Bernoulli random variables in $\{0, 1\}$, independent from $\mathcal{Z}$, with $\mathbb{P}(U_i = 0) = \mathbb{P}(U_i = 1) = 1/2$, and, for every $n \in \mathbb{N}$, let $S_n = (\mathcal{Z}_{Uj,j})_{1 \leq j \leq n}$.

Recall that $\mathrm{CMI}_\mu(\mathcal{A}) = I(\mathcal{A}(S); U|\mathcal{Z})$. Note that as was observed in [1], this quantity is equivalent to $I(\mathcal{A}(S); S|\mathcal{Z})$ when μ is atomless.

10 CMI of VC classes with finite star number

Recently it has been proven the optimal excess risk bound is also yielded by the CMI approach for stable compression schemes, such as SVM[15].

Theorem 20. Let $\mathcal{H}$ be a concept class with a stable compression scheme (κ, ρ) of size k. Then, for every realizable data distribution μ and $n \geq k$, $\mathrm{CMI}_\mu(\mathcal{A}_n) \leq 2k\log 2$, where $\mathcal{A}_n = \rho(\kappa(.))$.

References

[1] Mahdi Haghifam, Gintare Karolina Dziugaite, Shay Moran, and Daniel M. Roy. On the information complexity of proper learners for vc classes in the realizable case, 2020.

[2] Afshin Rostamizadeh Mehryar Mohri and Ameet Talwalkar. Foundations of machine learning, mit press, second edition, 2018.

[3] Thomas Steinke and Lydia Zakynthinou. Reasoning about generalization via conditional mutual information, 2020.

[4] Olivier Bousquet, Steve Hanneke, Shay Moran, and Nikita Zhivotovskiy. Proper learning, helly number, and an optimal svm bound, 2020.

[5] Steve Hanneke. The optimal sample complexity of pac learning, 2016.

[6] Andrzej Ehrenfeucht and David Haussler. A general lower bound on the number of examples needed for learning. 1989.

[7] Steve Hanneke. Refined error bounds for several learning algorithms, 2016.

[8] V. Vapnik and A. Chervonenkis. 1974.

[9] Steve Hanneke and Liu Yang. Minimax analysis of active learning, 2014.

[10] Aolin Xu and Maxim Raginsky. Information-theoretic analysis of generalization capability of learning algorithms, 2017.

[11] Daniel Russo and James Zou. Controlling bias in adaptive data analysis using information theory. 2016.

[12] R.M Dudley. The sizes of compact subsets of hilbert space and continuity of gaussian processes. pages 290–330, 1967.

[13] Amir R. Asadi, Emmanuel Abbe, and Sergio Verdú. Chaining mutual information and tightening generalization bounds, 2019.

[14] Hassan Hafez-Kolahi, Zeinab Golgooni, Shohreh Kasaei, and Mahdieh Soleymani. Conditioning and processing: Techniques to improve information-theoretic generalization bounds. 2020.

[15] Shay Moran Dan Roy Mahdi Haghifam, Gintare Karolina Dziugaite. Towards a unified information-theoretic framework for generalization. 2021.

[16] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. 1989.

A Proof of theorem 2

We prove the following theorem:

Theorem 21. Let G be a family of functions mapping from Z to $[0,1]$. Then, for any $\delta > 0$, with probability at least $1-\delta$, each of the following holds for all $g \in G$:

$$\mathbb{E}[g(z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(z_i) + 2\ \mathbb{E}[\hat{\mathcal{R}}_S(G)] + \sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \tag{2}$$

$$\mathbb{E}[g(z)] \leq \frac{1}{n}\sum_{i=1}^{n} g(z_i) + 2\ \hat{\mathcal{R}}_S(G) + 3\sqrt{\frac{\log(\frac{2}{\delta})}{2n}} \tag{3}$$

Proof. Denoting $E_S[g] = \frac{1}{n}\sum_{i=1}^{n} g(Z_i)]$, the proof revolves around applying McDiarmid's inequality to:

$$\Phi(S) = sup_{g\in G}\mathbb{E}[g] - E_S(g)$$

For two neighbour samples S, S' differing say in z_n,

$$\Phi(S) - \Phi(S') \leq sup_{g\in G} E_S(g) - E_{S'}(g) \leq \frac{1}{n}$$

Symmetrically we get that $|\Phi(S) - \Phi(S')| \leq \frac{1}{n}$, by McDiarmid's inequality with probability $1-\delta/2$:

$$\Phi(S) \leq \mathbb{E}[\Phi(S)] + \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$$

The part left to bound is $\mathbb{E}_S[\Phi(S)]$:

$$\begin{aligned}
\mathbb{E}_S[\Phi(S)] &= \mathbb{E}_S[sup_{g\in G}\ \mathbb{E}_{S'}[g] - E_S(g)] && (4)\\
&= \mathbb{E}_S[sup_{g\in G}\ \mathbb{E}_{S'}[E_{S'}(g) - E_S(g)]] && (5)\\
&\leq \mathbb{E}_{S,S'}[sup_{g\in G}\ E_{S'}(g) - E_S(g)] && (6)\\
&= \mathbb{E}_{S,S',\sigma}[sup_{g\in G}\ \frac{1}{n}\sum_{i=1}^{n}\sigma_i(g(Z_i) - g(Z_i'))] && (7)\\
&\leq 2\ \mathbb{E}_S[\hat{\mathcal{R}}_S(G)] && (8)
\end{aligned}$$

(6) comes from Jensen's inequality applied to the convex supremum function.
To get (4), we can use McDiarmid's inequality to $\hat{\mathcal{R}}_S(G)$, after observing that for two neighbors S, S', $|\hat{\mathcal{R}}_S(G) - \hat{\mathcal{R}}_{S'}(G)| \leq \frac{1}{n}$, which yields : $\mathbb{E}[\hat{\mathcal{R}}_S(G)] \leq \hat{\mathcal{R}}_S(G) + \sqrt{\frac{\log(\frac{2}{\delta})}{2n}}$. □

B Proof of theorem 4

Theorem 4 is a simple consequence of the following theorem:

Theorem 22. Let μ be a distribution on Z. Let $\mathcal{A}$ be a randomized algorithm. Let $\ell : \times Z \to \mathbb{R}$ be an arbitrary (deterministic, measurable) function. Suppose there exists $\Delta : Z^2 \to \mathbb{R}$ such that $|\ell(h, z_1) - \ell(h, z_2)| \leq \Delta(z_1, z_2)$ for all $z_1, z_2 \in Z$ and $h \in \mathcal{H}$. Then:

$$|gen_\mu(\mathcal{A})| \leq \sqrt{\frac{2 \text{ CMI}_\mu(\mathcal{A})}{n} E_{\mu^{\otimes 2}}[\Delta(Z_1, Z_2)^2]}$$

Proof. The proof is mostly based on the following lemma:

Lemma 23. Let X and Y be random variables on Ω (with X absolutely continuous with respect to Y) and $f : \Omega \to \mathbb{R}$ a (measurable) function. Then

$$\mathbb{E}[f(X)] \leq D_{KL}(X||Y) + \log \mathbb{E}[e^{f(Y)}]$$

And on the following corollary:
Let S, S', and Z be independent random variables where S and S' have identical distributions. Let A be a random function independent from S, S', and Z. Let g be a fixed function. Then

$$\mathbb{E}_{A,S,Z}[g(A(S,Z),S,Z)] \leq \mathbb{E}_Z[\inf_{t>0} \frac{I(A(S,Z),S) + \log \mathbb{E}_{A,S',S,Z}[e^{tg(A(S,Z),S',Z)}]}{t}] \tag{9}$$

$$\leq \inf_{t>0} \frac{I(A(S,Z),S|Z) + \mathbb{E}_Z[\log \mathbb{E}_{A,S',S,Z}[e^{tg(A(S,Z),S',Z)}]]}{t} \tag{10}$$

□

C Proof of lemma 9

Proof.

- Recall $\mathcal{X}_{\mathcal{H}',l} = \{x \in \mathcal{X} : \sum_{h \in \mathcal{H}'} \mathbb{1}[h(x) \neq h_{maj}(x) < \frac{|\mathcal{H}'|}{l}]\}$.
  Given exactly k_o classifiers $\mathcal{H}'$ in $\mathcal{H}$ that are witnessing an hollow star sequence of points, for any $l < k_0$, the region $\mathcal{X}_{\mathcal{H}',l}$ must contain the hollow star sequence. Therefore the majority vote of the $\mathcal{H}'$ classifiers is unrealizable on $\mathcal{X}_{\mathcal{H}',l}$. So, $k_p \geq k_o$.

- Suppose $k_w < \infty$. If for some finite multiset $\mathcal{H}' \subseteq \mathcal{H}$ there is no $h \in \mathcal{H}$ that coincides with Majority($\mathcal{H}'$) on $\mathcal{X}_{\mathcal{H}',k_w}$, then $S = \{(x, \text{Majority}(\mathcal{H}')(x)) : x \in \mathcal{X}_{\mathcal{H}',k_w}\}$is an unrealizable set. Therefore, it contains a subset W of size at most k_w that is also unrealizable. By definition of $\mathcal{X}_{\mathcal{H}',l}$, each point (x, y) in W contradicts strictly fewer than $|\mathcal{H}'|/k_w$ elements of $\mathcal{H}'$, which contradicts the unrealizability of S. Therefore, $k_p \leq k_w$.

D Proof of the optimal complexity of improper learning

The proof uses the following lemma [16]:

Lemma 24. For any $\delta \in (0,1)$, $n \in \mathbb{N}$, $f^* \in \mathcal{H}$ and any probability measure μ over $\mathcal{X}$, letting $Z_1, \ldots, Z_n$ be independent μ-distributed random variables, with probability at least $1-\delta$, every $h \in \mathcal{H}[\{(Z_i, f^*(Z_i))\}_{i \leq n}]$ satisfies

$$er_\mu(h, f^*) = \mu(\{x : h(x) \neq f^*(x)\}) \leq \frac{2}{n}\left(d \log_2(\frac{2en}{d}) + \log_2(\frac{2}{\delta})\right)$$

The lemma gives an upper bound of the probability that two consistent learners on n samples disagree.
Now we fix $f^* \in \mathcal{H}$, a probability measure μ over $\mathcal{X}$, and for brevity use $\mathbb{S}_{1:m} = (\mathbb{Z}_{1:m}, f^*(\mathbb{Z}_{1:m}))$,
For any classifier h, define $\mathrm{ER}(h) := \{x \in \mathcal{X} : h(x) \neq f^*(x)\}$.
First notice that for any $S' \subseteq S$ and T, $\mathcal{H}[S \cup T] = \mathcal{H}[S] \cap \mathcal{H}[T]$ and that $\mathcal{H}[S] \subseteq \mathcal{H}[S']$. So by construction for any $S' \in \mathcal{A}(S,T)$, $\mathcal{H}[S'] \subseteq \mathcal{H}[T]$, hence the well definition of the objects that follow.

Claim 25. For any $n \in \mathbb{N}$, for every $\delta \in (0,1)$, and every finite sequence T of points in $\mathcal{X} \times \mathcal{Y}$ and $f^* \in \mathcal{H}[T]$, wp. at least $1-\delta$, then $h_{n,T} = \mathrm{Majority}(L(\mathcal{A}(\mathbb{S}_{1:n}; T)))$ satisfies

$$er_\mu(h_{n,T}, f^*) \leq \frac{c}{n+1}\left(d + \ln(\frac{18}{\delta})\right)$$

The claim yields the theorem by taking $T = \emptyset$, and $n \geq \frac{c}{\varepsilon}\left(d + \ln\frac{18}{\delta}\right)$.

Proof of claim 25: We prove the claim by induction on n. The base case is direct by taking $n < c\,\ln(18e) - 1$.

Suppose the claim holds for every $m < n$ integers, we want to show it holds for n as well, and in particular $n \geq 4$. Let S_0, S_1, S_2, S_3 be as in the definition of $\mathcal{A}(S;T)$, with $S = S_{1:n}$. Also denote $T_1 = S_2 \cup S_3 \cup T, T_2 \ldots$ and $h_i = \mathrm{Majority}(L(A(S_0; T_i)))$. It is easy to check that the h_i are well defined. By inductive hypothesis, setting $\delta' = \delta/9$, we get for an event E_1^i of probability at least $1 - \delta/9$:

$$\mu(\mathrm{ER}(h_i)) \leq \frac{c}{|S_0|+1}\left(d + \ln(\frac{9 \cdot 18}{\delta})\right) \leq \frac{4c}{n}\left(d + \ln(\frac{9 \cdot 18}{\delta})\right) \tag{11}$$

We denote by R_i the sequence of elements in $S_i \cap (\mathrm{ER}(h_i) \times \mathcal{Y}) = \{(Z_{i,t}, f^*(Zi,t))\}_{t \leq N_i}$. Since h_i and S_i are independent, $Z_{i,1}, \ldots, Z_{i,N_i}$ are conditionally independent given h_i and N_i, each with conditional distribution $\mu(.|\mathrm{ER}(h_i)$. Thus, applying lemma 24 on probability at least $1 - \delta/9$, if $N_i > 0$, every $h \in \mathcal{H}[R_i]$ for an event E_2^i of probability at least $1 - \delta/9$:

$$er_{\mu(.|\mathrm{ER}(h_i))}(h, f^*) \leq \frac{2}{N_i}\left(d \log_2(\frac{2eN_i}{d}) + \log_2(\frac{18}{\delta})\right)$$

Since for $j \neq i$ we have $\mathcal{H}[T_j] \subseteq \mathcal{H}[R_i]$, then for $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L(\mathcal{A}(S_0, T_j))$, we get for E_2^i:

$$\begin{aligned}\mu(\mathrm{ER}(h_i) \cap \mathrm{ER}(h)) &= \mu(\mathrm{ER}(h)|\mathrm{ER}(h_i))\mu(\mathrm{ER}(h_i)) \\ &\leq \mu(\mathrm{ER}(h_i))\frac{2}{N_i}\left(d\,\log_2(\frac{2eN_i}{d}) + \log_2(\frac{18}{\delta})\right) \end{aligned} \tag{12}$$

Since h_i and S_i are independent, by Chernoff bound (applied under the conditional distribution given h_i) and the law of total probability, then there is an event E_3^i of probability at least $1 - \delta/9$ for which $\mu(\mathrm{ER}(h_i)) \geq \frac{23}{\lfloor n/4 \rfloor}\ln(\frac{9}{\delta}) \geq \frac{2(10/3)^2}{\lfloor n/4 \rfloor}\ln(\frac{9}{\delta})$ and:

$$N_i \geq \frac{7|S_i|}{10}\mu(\mathrm{ER}(h_i)) = \frac{7n}{40}\mu(\mathrm{ER}(h_i)) \tag{13}$$

Now on $E_1^i \cap E_2^i \cap E_3^i$, if $\mu(\mathrm{ER}(h_i)) \geq \frac{23}{\lfloor n/4 \rfloor}\ln(\frac{9}{\delta})$ we have that $N_i > 0$ and so for every $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L(\mathcal{A}(S_0, T_j))$ combining (11),(12),(13):

$$\begin{aligned}\mu(\mathrm{ER}(h_i) \cap \mathrm{ER}(h)) &\leq \mu(\mathrm{ER}(h_i))\frac{2}{N_i}\left(d\,\log_2(\frac{2eN_i}{d}) + \log_2(\frac{18}{\delta})\right) \\ &\leq \frac{20}{7\ln\,2(\lfloor n/4 \rfloor)}\left(d\,\log(\frac{2e\frac{7n}{40}\mu(\mathrm{ER}(h_i))}{d}) + \log(\frac{18}{\delta})\right) \\ &\leq \frac{20}{7\ln\,2(\lfloor n/4 \rfloor)}\left(d\,\log(\frac{2e\frac{7n}{40}\frac{4c}{n}\left(d + \ln(\frac{9\cdot 18}{\delta})\right)}{d}) + \log(\frac{18}{\delta})\right) \\ &\leq \frac{20}{7\ln\,2(\lfloor n/4 \rfloor)}\left(d\,\ln(\frac{9ec}{5}) + \ln\,(\frac{18}{\delta})\right) \end{aligned} \tag{14}$$

(14) uses the fact that $n > c\,\ln(18e) - 1 > 3200$ and $\lfloor n/4 \rfloor > (n-4)/4 > \frac{799}{800}\frac{3200}{3201}\frac{n+1}{4}$. (14) is also less than $\frac{150}{n+1}\left(d + \ln\,(\frac{18}{\delta})\right)$.
Notice that if $\mu(\mathrm{ER}(h_i)) < \frac{23}{\lfloor n/4 \rfloor}\ln(\frac{9}{\delta})$, then $\mu(\mathrm{ER}(h_i) \cap \mathrm{ER}(h)) \leq \mu(\mathrm{ER}(h_i)) < \frac{23}{\lfloor n/4 \rfloor}\ln(\frac{9}{\delta}) < \frac{150}{n+1}\left(d + \ln\,(\frac{18}{\delta})\right)$.

Therefore on $E_1^i \cap E_2^i \cap E_3^i$, for every $h \in \bigcup_{j\in\{1,2,3\}\setminus\{i\}} L(\mathcal{A}(S_0, T_j))$:

$$\mu(\mathrm{ER}(h_i) \cap \mathrm{ER}(h)) < \frac{150}{n+1}\left(d + \ln\,(\frac{18}{\delta})\right)$$

Now consider $h_{maj} = \mathrm{Majority}(L(\mathcal{A}(S;T)))$ with $S = \mathbb{S}_{1:n}$. By pigeonhole principle for $x \in \mathcal{X}$, there is i such that $h_i(x) = h_{maj}(x)$. Since $\{L(\mathcal{A}(S_0;T_i))\}_i$ 3-equipartitions $L(\mathcal{A}(S;T)$, it must be that $\bigcup_{j\in\{1,2,3\}\setminus\{i\}} L(\mathcal{A}(S_0;T_j))$ has $1/4$ of its classifiers agreeing with h_{maj} on any $x \in \mathcal{X}$.
Therefore for I a random variable uniformly distributed on $\{1,2,3\}$ (independent of the data), $\tilde{h}$ a random variable conditionally (given I and S) uniformly distributed on the classifiers $\bigcup_{j\{1,2,3\}\setminus\{I\}} L(\mathcal{A}(S_0;T_j))$ for any fixed $x1 \in \mathrm{ER}(h_{maj})$, with conditional (given S) probability at least $1/12$, $h_I(x) = \tilde{h}(x) = h_{maj}(x)$, so that $x \in \mathrm{ER}(h_I) \cap \mathrm{ER}(\tilde{h})$ as well. So for X a random variable of distribution μ independent of I and $\tilde{h}$:

$$\mathbb{E}\left[\mu(\mathrm{ER}(h_I)\cap\mathrm{ER}(\tilde{h}))|S\right] = \mathbb{E}\left[\mathbb{P}(X\in\mathrm{ER}(h_I)\cap\mathrm{ER}(\tilde{h})|I,\tilde{h},S)|S\right]$$
$$= \mathbb{E}\left[\mathbb{P}(X\in\mathrm{ER}(h_I)\cap\mathrm{ER}(\tilde{h})|\tilde{h},S)|S\right]$$
$$\geq \mathbb{E}\left[\mathbb{P}(X\in\mathrm{ER}(h_I)\cap\mathrm{ER}(\tilde{h})|\tilde{h},S)1_{X\in\mathrm{ER}(h_{maj})}|S\right]$$
$$\geq \mathbb{E}\left[1/121_{X\in\mathrm{ER}(h_{maj})}|S\right] = 1/12\ er_\mu(h_{maj},f^*)$$

Finally on $\bigcap_i E_1^i\cap E_2^i\cap E_3^i$:

$$er_\mu(h_{maj},f^*) \leq 12\ \mathbb{E}\left[\mu(\mathrm{ER}(h_I)\cap\mathrm{ER}(\tilde{h}))|S\right]$$
$$\leq 12\ max_i max_{j\neq i} max_{h\in L(\mathcal{A}(S_0;T_j))}\mu(\mathrm{ER}(h_i)\cap\mathrm{ER}(h))$$
$$\leq \frac{1800}{n+1}(d+\ln\,(\frac{18}{\delta})) = \frac{c}{n+1}(d+\ln\,(\frac{18}{\delta}))$$

By union bound the event $\bigcap_i E_1^i\cap E_2^i\cap E_3^i$ has probability at least $1-\delta$ for any $\delta\in(0,1)$. Hence the claim by induction.

□

□