30 Compression Bounds

Throughout the book, we have tried to characterize the notion of learnability using different approaches. At first we have shown that the uniform convergence property of a hypothesis class guarantees successful learning. Later on we introduced the notion of stability and have shown that stable algorithms are guaranteed to be good learners. Yet there are other properties which may be sufficient for learning, and in this chapter and its sequel we will introduce two approaches to this issue: compression bounds and the PAC-Bayes approach.

In this chapter we study compression bounds. Roughly speaking, we shall see that if a learning algorithm can express the output hypothesis using a small subset of the training set, then the error of the hypothesis on the rest of the examples estimates its true error. In other words, an algorithm that can "compress" its output is a good learner.

30.1 Compression Bounds

To motivate the results, let us first consider the following learning protocol. First, we sample a sequence of k examples denoted T. On the basis of these examples, we construct a hypothesis denoted h_T. Now we would like to estimate the performance of h_T so we sample a fresh sequence of $m-k$ examples, denoted V, and calculate the error of h_T on V. Since V and T are independent, we immediately get the following from Bernstein's inequality (see Lemma B.10).

LEMMA 30.1 *Assume that the range of the loss function is* $[0,1]$. *Then,*

$$\mathbb{P}\left[L_{\mathcal{D}}(h_T) - L_V(h_T) \geq \sqrt{\frac{2L_V(h_T)\log(1/\delta)}{|V|}} + \frac{4\log(1/\delta)}{|V|}\right] \leq \delta.$$

To derive this bound, all we needed was independence between T and V. Therefore, we can redefine the protocol as follows. First, we agree on a sequence of k indices $I = (i_1, \ldots, i_k) \in [m]^k$. Then, we sample a sequence of m examples $S = (z_1, \ldots, z_m)$. Now, define $T = S_I = (z_{i_1}, \ldots, z_{i_k})$ and define V to be the rest of the examples in S. Note that this protocol is equivalent to the protocol we defined before – hence Lemma 30.1 still holds.

Applying a union bound over the choice of the sequence of indices we obtain the following theorem.

THEOREM 30.2 *Let k be an integer and let $B : Z^k \to \mathcal{H}$ be a mapping from sequences of k examples to the hypothesis class. Let $m \geq 2k$ be a training set size and let $A : Z^m \to \mathcal{H}$ be a learning rule that receives a training sequence S of size m and returns a hypothesis such that $A(S) = B(z_{i_1}, \ldots, z_{i_k})$ for some $(i_1, \ldots, i_k) \in [m]^k$. Let $V = \{z_j : j \notin (i_1, \ldots, i_k)\}$ be the set of examples which were not selected for defining $A(S)$. Then, with probability of at least $1 - \delta$ over the choice of S we have*

$$L_{\mathcal{D}}(A(S)) \;\leq\; L_V(A(S)) + \sqrt{L_V(A(S))\frac{4k\log(m/\delta)}{m}} + \frac{8k\log(m/\delta)}{m}.$$

Proof For any $I \in [m]^k$ let $h_I = B(z_{i_1}, \ldots, z_{i_k})$. Let $n = m - k$. Combining Lemma 30.1 with the union bound we have

$$\begin{aligned}
&\mathbb{P}\left[\exists I \in [m]^k \text{ s.t. } L_{\mathcal{D}}(h_I) - L_V(h_I) \geq \sqrt{\frac{2L_V(h_I)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}\right] \\
&\quad\leq \sum_{I\in[m]^k} \mathbb{P}\left[L_{\mathcal{D}}(h_I) - L_V(h_I) \geq \sqrt{\frac{2L_V(h_I)\log(1/\delta)}{n}} + \frac{4\log(1/\delta)}{n}\right] \\
&\quad\leq m^k\delta.
\end{aligned}$$

Denote $\delta' = m^k\delta$. Using the assumption $k \leq m/2$, which implies that $n = m - k \geq m/2$, the above implies that with probability of at least $1 - \delta'$ we have that

$$L_{\mathcal{D}}(A(S)) \;\leq\; L_V(A(S)) + \sqrt{L_V(A(S))\frac{4k\log(m/\delta')}{m}} + \frac{8k\log(m/\delta')}{m},$$

which concludes our proof. □

As a direct corollary we obtain:

COROLLARY 30.3 *Assuming the conditions of Theorem 30.2, and further assuming that $L_V(A(S)) = 0$, then, with probability of at least $1-\delta$ over the choice of S we have*

$$L_{\mathcal{D}}(A(S)) \;\leq\; \frac{8k\log(m/\delta)}{m}.$$

These results motivate the following definition:

DEFINITION 30.4 **(Compression Scheme)** Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\mathcal{Y}$ and let k be an integer. We say that $\mathcal{H}$ has a compression scheme of size k if the following holds:
For all m there exists $A : Z^m \to [m]^k$ and $B : Z^k \to \mathcal{H}$ such that for all $h \in \mathcal{H}$, if we feed any training set of the form $(x_1, h(x_1)), \ldots, (x_m, h(x_m))$ into A and then feed $(x_{i_1}, h(x_{i_1})), \ldots, (x_{i_k}, h(x_{i_k}))$ into B, where $(i_1, \ldots, i_k)$ is the output of A, then the output of B, denoted h', satisfies $L_S(h') = 0$.

It is possible to generalize the definition for unrealizable sequences as follows.

DEFINITION 30.5 **(Compression Scheme for Unrealizable Sequences)** Let $\mathcal{H}$ be a hypothesis class of functions from $\mathcal{X}$ to $\mathcal{Y}$ and let k be an integer. We say that $\mathcal{H}$ has a compression scheme of size k if the following holds:
For all m there exists $A : Z^m \to [m]^k$ and $B : Z^k \to \mathcal{H}$ such that for all $h \in \mathcal{H}$, if we feed any training set of the form $(x_1, y_1), \ldots, (x_m, y_m)$ into A and then feed $(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$ into B, where $(i_1, \ldots, i_k)$ is the output of A, then the output of B, denoted h', satisfies $L_S(h') \leq L_S(h)$.

The following lemma shows that the existence of a compression scheme for the realizable case also implies the existence of a compression scheme for the unrealizable case.

LEMMA 30.6 *Let $\mathcal{H}$ be a hypothesis class for binary classification, and assume it has a compression scheme of size k in the realizable case. Then, it has a compression scheme of size k for the unrealizable case as well.*

Proof Consider the following scheme: First, find an ERM hypothesis and denote it by h. Then, discard all the examples on which h errs. Now, apply the realizable compression scheme on the examples that have not been removed. The output of the realizable compression scheme, denoted h', must be correct on the examples that have not been removed. Since h errs on the removed examples it follows that the error of h' cannot be larger than the error of h; hence h' is also an ERM hypothesis. □

30.2 Examples

In the examples that follows, we present compression schemes for several hypothesis classes for binary classification. In light of Lemma 30.6 we focus on the realizable case. Therefore, to show that a certain hypothesis class has a compression scheme, it is necessary to show that there exist A, B, and k for which $L_S(h') = 0$.

30.2.1 Axis Aligned Rectangles

Note that this is an uncountable infinite class. We show that there is a simple compression scheme. Consider the algorithm A that works as follows: For each dimension, choose the two positive examples with extremal values at this dimension. Define B to be the function that returns the minimal enclosing rectangle. Then, for $k = 2d$, we have that in the realizable case, $L_S(B(A(S))) = 0$.

30.2.2 Halfspaces

Let $\mathcal{X} = \mathbb{R}^d$ and consider the class of homogenous halfspaces, $\{\mathbf{x} \mapsto \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle) : \mathbf{w} \in \mathbb{R}^d\}$.

A Compression Scheme:
W.l.o.g. assume all labels are positive (otherwise, replace $\mathbf{x}_i$ by $y_i\mathbf{x}_i$). The compression scheme we propose is as follows. First, A finds the vector $\mathbf{w}$ which is in the convex hull of $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ and has minimal norm. Then, it represents it as a convex combination of d points in the sample (it will be shown later that this is always possible). The output of A are these d points. The algorithm B receives these d points and set $\mathbf{w}$ to be the point in their convex hull of minimal norm.

Next we prove that this indeed is a compression sceme. Since the data is linearly separable, the convex hull of $\{\mathbf{x}_1, \ldots, \mathbf{x}_m\}$ does not contain the origin. Consider the point $\mathbf{w}$ in this convex hull closest to the origin. (This is a unique point which is the Euclidean projection of the origin onto this convex hull.) We claim that $\mathbf{w}$ separates the data.[1] To see this, assume by contradiction that $\langle \mathbf{w}, \mathbf{x}_i \rangle \leq 0$ for some i. Take $\mathbf{w}' = (1-\alpha)\mathbf{w} + \alpha\mathbf{x}_i$ for $\alpha = \frac{\|\mathbf{w}\|^2}{\|\mathbf{x}_i\|^2 + \|\mathbf{w}\|^2} \in (0,1)$. Then $\mathbf{w}'$ is also in the convex hull and

$$\begin{aligned}
\|\mathbf{w}'\|^2 &= (1-\alpha)^2\|\mathbf{w}\|^2 + \alpha^2\|\mathbf{x}_i\|^2 + 2\alpha(1-\alpha)\langle \mathbf{w}, \mathbf{x}_i \rangle \\
&\leq (1-\alpha)^2\|\mathbf{w}\|^2 + \alpha^2\|\mathbf{x}_i\|^2 \\
&= \frac{\|\mathbf{x}_i\|^4\|\mathbf{w}\|^2 + \|\mathbf{x}_i\|^2\|\mathbf{w}\|^4}{(\|\mathbf{w}\|^2 + \|\mathbf{x}_i\|^2)^2} \\
&= \frac{\|\mathbf{x}_i\|^2\|\mathbf{w}\|^2}{\|\mathbf{w}\|^2 + \|\mathbf{x}_i\|^2} \\
&= \|\mathbf{w}\|^2 \cdot \frac{1}{\|\mathbf{w}\|^2/\|\mathbf{x}_i\|^2 + 1} \\
&< \|\mathbf{w}\|^2,
\end{aligned}$$

which leads to a contradiction.

We have thus shown that $\mathbf{w}$ is also an ERM. Finally, since $\mathbf{w}$ is in the convex hull of the examples, we can apply Caratheodory's theorem to obtain that $\mathbf{w}$ is also in the convex hull of a subset of $d+1$ points of the polygon. Furthermore, the minimality of $\mathbf{w}$ implies that $\mathbf{w}$ must be on a face of the polygon and this implies it can be represented as a convex combination of d points.

It remains to show that $\mathbf{w}$ is also the projection onto the polygon defined by the d points. But this must be true: On one hand, the smaller polygon is a subset of the larger one; hence the projection onto the smaller cannot be smaller in norm. On the other hand, $\mathbf{w}$ itself is a valid solution. The uniqueness of projection concludes our proof.

30.2.3 Separating Polynomials

Let $\mathcal{X} = \mathbb{R}^d$ and consider the class $\mathbf{x} \mapsto \text{sign}(p(x))$ where p is a degree r polynomial.

[1] It can be shown that $\mathbf{w}$ is the direction of the max-margin solution.

Note that $p(x)$ can be rewritten as $\langle \mathbf{w}, \psi(\mathbf{x}) \rangle$ where the elements of $\psi(x)$ are all the monomials of $\mathbf{x}$ up to degree r. Therefore, the problem of constructing a compression scheme for $p(\mathbf{x})$ reduces to the problem of constructing a compression scheme for halfspaces in $\mathbb{R}^{d'}$ where $d' = O(d^r)$.

30.2.4 Separation with Margin

Suppose that a training set is separated with margin γ. The Perceptron algorithm guarantees to make at most $1/\gamma^2$ updates before converging to a solution that makes no mistakes on the entire training set. Hence, we have a compression scheme of size $k \leq 1/\gamma^2$.

30.3 Bibliographic Remarks

Compression schemes and their relation to learning were introduced by Littlestone & Warmuth (1986). As we have shown, if a class has a compression scheme then it is learnable. For binary classification problems, it follows from the fundamental theorem of learning that the class has a finite VC dimension. The other direction, namely, whether every hypothesis class of finite VC dimension has a compression scheme of finite size, is an open problem posed by Manfred Warmuth and is still open (see also (Floyd 1989, Floyd & Warmuth 1995, Ben-David & Litman 1998, Livni & Simon 2013).