

# Generalization and the VC Dimension

## 1 Overview

So far we have studied the learning problem. The general approach is as follows. Select a learning model. Usually such learning models will have “complexity parameters” and “tuning parameters”. For example, for the neural network learning model, the number of hidden layers and the number of hidden units in each layer would quantify the complexity of the neural network. Two properties of a “good” learning model are that

1. It should be possible on any data set to guarantee a training error  $R_{emp}$  of zero by tuning the complexity parameters. This is a universal approximation requirement. For example, if I tell you that you will be given 500 data points, you can then go and choose a network architecture for which you know a training error of zero is possible.
2. A good (computationally efficient) learning algorithm exists for selecting the tuning parameters. Usually the tuning parameters are selected so as to minimize  $R_{emp}$ .

We have seen that neural networks satisfy these criteria. The multi-layer perceptron has the universal approximation property and for learning algorithms that are based on minimizing  $E(\mathbf{w})$  where  $E(\mathbf{w})$  is a differentiable function of the weights, a computationally adequate iterative algorithm exists – gradient descent – with the gradients being obtained via back-propagation.

Our ultimate goal is to minimize the the test error  $R$ . Thus, an important question to address is whether minimizing  $R_{emp}$  is equivalent to minimizing  $R$ , at least to within some degree of accuracy  $\epsilon$ . This is the topic of generalization.

## 2 Coin/Bin Model

Imagine that every function is a coin. For a given function  $g(\mathbf{x})$ , the probability that the coin pops heads is analogous to the probability of error, i.e., the probability that  $g(\mathbf{x}) \neq f(\mathbf{x})$ . Drawing a data set of size  $N$  and then computing the training error for the particular function is then analogous to flipping that particular coin  $N$  times. The complete analogy is given in the next table.

	Learning Problem	Coin/Bin Model
<b>Targen Function</b>	$f(\mathbf{x})$	
<b>Learning Model</b>	$g_1(\mathbf{x}), g_2(\mathbf{x}), \dots$	$coin_1, coin_2, \dots$
<b>Test error of function <math>i</math></b>	$P[g_i(\mathbf{x}) \neq f(\mathbf{x})]$	$\pi_i = P[coin_i \text{ pops heads}]$
<b>Draw an <i>iid</i> data set of size <math>N</math></b>	Randomly select $N$ $x$ 's	Flip each coin $N$ times
<b>Compute training error</b>	fraction of disagreements, $R_{emp}$	fraction of heads, $\nu$
<b>Learning algorithm</b>	minimize $R_{emp}$	Pick coin with lowest $\nu$
<b>Generalization error for function <math>i</math></b>	$ R - R_{emp} $ of function $i$	$ \pi_i - \nu_i $

## 3 Single Function Learning Model

A learning model with a single function in it is analogous to a single coin. Let the probability of heads be  $\pi$ . When a data set is drawn, the probability that the training error is  $\nu = \frac{k}{N}$  is given by the binomial distribution

$$P[\nu|\pi] = \binom{N}{N\nu} \pi^{N\nu} (1 - \pi)^{N(1-\nu)} \quad (1)$$

The appropriate generalization question to ask is, how likely is it that the observed  $\nu$  deviates from  $\pi$  by a given tolerance  $\epsilon$ . This is formalized in the quantity  $P[|\nu - \pi| \geq \epsilon|\pi]$ . If we fix  $\epsilon$ , ideally this quantity should

be small. Thus, it is useful to have a bound for this quantity, and many bounds exist, the most useful one being due to Chernoff which gives that

$$P[|\nu - \pi| \geq \epsilon|\pi] \leq 2e^{-2N\epsilon^2} \quad (2)$$

There are many interesting points to be noticed.

1. This bound is independent of  $\nu$  and  $\pi$ . Thus, this bound holds for *any* target function, *any* single function  $g$  and *any* input distribution. Note that these are the three things that determine  $\pi$  and  $\nu$ .
2. For any fixed  $\epsilon$ , the bound approaches zero very rapidly as  $N$  increases. Thus to within any pre-specified tolerance, one has good generalization as  $N \rightarrow \infty$ .

A useful rule of thumb can be deduced from this result by turning it around. (2) says that the probability of a large deviation is small. Equivalently, the probability of a small deviation is large. More formally, letting  $\eta = 2e^{-2N\epsilon^2}$

$$P[|\nu - \pi| < \epsilon|\pi] \geq 1 - \eta \quad (3)$$

or, with probability greater than  $1 - \eta$ ,

$$\pi < \nu + \sqrt{\frac{\log(2/\eta)}{2N}} \quad (4)$$

in other words, with very high probability,  $\pi$  is at most  $\nu$  plus a “small” (decreasing like  $1/\sqrt{N}$ ) quantity. Choosing  $\log(2/\eta) \approx 1$  we have the rule of thumb that if you want generalization with accuracy  $\epsilon$  then you need roughly  $1/\epsilon^2$  training examples.

## 4 Finitely Many Functions

We now consider the case that there are finitely many functions,  $M$  of them  $(g_1, \dots, g_M)$ . Since we do not wish to restrict ourselves to a particular learning algorithm, we want it to be the case that no matter which coin the learning algorithm ends up with, the generalization will be good. In particular, if the learning algorithm simply picks a coin at random, then we are effectively back to the case of a single function. On the other hand, the learning algorithm could pick a function based upon its  $\nu$  causing the resulting  $\nu$  to be biased in some way. For example the learning algorithm could pick the function with the lowest  $\nu$ . In this case the quantity of interest would be  $P[|\nu_{min} - \pi_{min}| \geq \epsilon|\pi_1, \dots, \pi_M]$ . Without knowing what the learning algorithm is, how can one be sure that the function you end up with has good generalization. The only way is to ensure that  $|\nu - \pi|$  is small for every function. One will then achieve good generalization independent of the learning algorithm. Thus, the quantity we wish to bound is  $P[\max_i |\nu_i - \pi_i| \geq \epsilon|\pi_1, \dots, \pi_M]$ . From now on we will drop the conditioning on the  $\pi_i$ 's, understanding that it is there. Since this is equivalent to  $P[|\nu_1 - \pi_1| \geq \epsilon \text{ OR } |\nu_2 - \pi_2| \geq \epsilon, \dots, \text{ OR } |\nu_M - \pi_M| \geq \epsilon]$ , an application of the union bound gives that

$$P[\max_i |\nu_i - \pi_i| \geq \epsilon] \leq 2Me^{-2N\epsilon^2} \quad (5)$$

Once again, this bound is independent of  $\nu_i$ ,  $\pi_i$  and thus is independent of the target function, the particular nature of the learning model (as long as it has finitely many functions), and the input distribution. A similar analysis to the single function case yields that with probability greater than  $1 - \eta$ ,

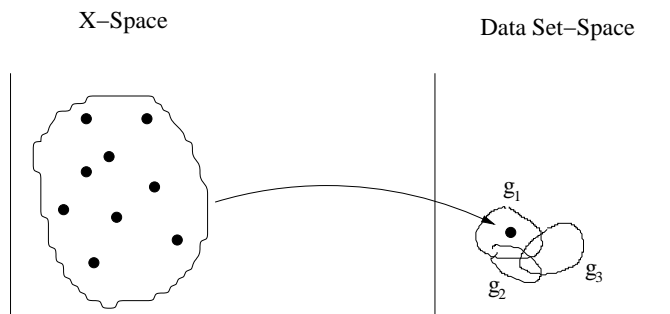
$$\pi < \nu + \sqrt{\frac{\log(2M/\eta)}{2N}} \quad (6)$$

## 5 Infinitely Many Functions

Unfortunately, we cannot simply let  $M \rightarrow \infty$  in (5) above. For starters, the limit does not converge on the right hand side of (5). The left hand side is also not well defined, and the “infinite” version of (5) would be given by a quantity of the form

$$P \left[ \sup_{g \in \mathcal{L}} |\nu_g - \pi_g| \geq \epsilon \right] \leq (?) \quad (7)$$

where it is now necessary to take a sup over functions in the learning model as the max is not a well defined operation for infinitely many quantities. To understand how we might be able to get a bound for infinitely many functions, first consider the case that there are only  $M$  functions,  $g_1, g_2, g_2, \dots, g_2$ . That is  $g_2$  is repeated in our learning model  $M - 1$  times. Effectively there are only 2 functions in our learning model, and though the bound in (5) is valid, it is very loose.  $M$  could be replaced by 2. In general, we could replace  $M$  by  $M_{eff}$ , the number of different functions. What if the repeated  $g_2$  in the learning model, are replaced by different but very similar functions to  $g_2$ . The bound (5) is still true but we suspect that there might be a huge slack and we suspect that there might be some  $M_{eff} < M$  for which (5) will hold, where  $M_{eff}$  will depend on how similar the functions are. In this case we might expect  $M_{eff}$  to be only slightly larger than 2. The hope is that while  $M \rightarrow \infty$ , it might be possible that  $M_{eff}$  still remains finite and can be used in a bound like (5). An illustration of why this might happen is given in the following figure.



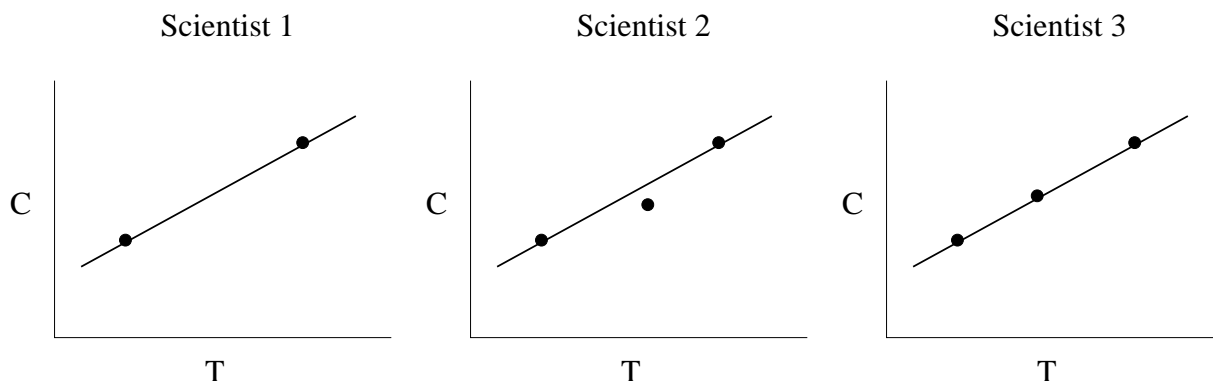
In the  $X$  space, a data set is represented by a set of points. Imagine an abstract space, the data set space, where this data set is represented by a single point. For example if the data set consists of  $N$  points in  $\mathbf{R}^2$ , this data set can be represented by a point in  $\mathbf{R}^{2N}$ . A region in data set space corresponds to a collection of different data sets. For a given data set (point in data set space), either  $|\nu_1 - \pi_1| \geq \epsilon$  or not. If it is, then we will say that the data set is bad. The set of all bad data sets for  $g_1$ , defines a region in data set space and is illustrated by the region  $g_1$  in the figure. Similarly the regions  $g_2$ , and  $g_3$ . According to our criterion, a “bad” event occurs if a dataset falls in any one of the regions. Thus we are interested in the probability that a data set falls within the union of these regions. The “volume” of each region is bounded by the Chernoff inequality. The volume of the union is bounded by the sum of the individual volumes (union bound). But, if the functions are very similar, these regions will be nearly identical and the intersections will be huge. Thus a tighter bound may be possible (perhaps, say, as a function of the minimum size of an intersection). If on the other hand the functions are very dissimilar, then the intersections would be small and we might be out of luck, and we may essentially not be able to improve on the union bound - in this case perhaps we cannot get good generalization under our strict criterion.

### 5.1 Axiom of Non-Falsifiability

In order to get a handle on how to proceed in obtaining this tighter bound, and a feel for the abstract quantities that we will be studying, our initial approach will be through the axiom of non-falsifiability, a general principle in data driven sciences. First some illustrations.

1. Suppose that one constructs a physical theory about the conductivity of a metal under various temperatures. In this theory, aside for some constants that need to be determined, the conductivity  $C$  has

a linear dependence on the temperature  $T$ . In order to verify that the theory is correct and to obtain the unknown constants, 3 scientists conduct the following three experiments and present their data to you.



It is clear that Scientist 3 has produced the most convincing evidence for the theory. If the measurements are exact, then, Scientist 2 has managed to falsify the theory and we are back to the drawing board. What about Scientist 1. While he has not falsified the theory, has he provided any evidence for it? The answer is no, for we can reverse the question. Suppose that the theory was not correct, what could the data have done to prove him wrong - nothing. The theory is not falsifiable under the conditions of experiment 1 and so the first experiment cannot be viewed as providing evidence for or against the theory.

2. Financial firms try to pick good traders (predictors of whether the market will go up or not). Suppose that each trader is tested on their prediction (up or down) over the next 5 days and those who perform well will be hired. One might think that this process should produce better and better traders on wall street. Viewed as a learning problem, consider each trader as a prediction function. Suppose that there are  $2^5$  traders who happen to be a diverse set of people such that their predictions over the next 5 days are all different. Necessarily one of these traders gets it all correct, and will be hired. Hiring the trader may or may not be a good thing, but this experiment cannot determine it. Why, because the hypothesis, “a perfect predictor exists in this group” is not falsifiable under this experiment.

**Axiom 5.1 (Axiom of Non-Falsifiability)** *If the outcome of an experiment can not falsify a particular hypothesis, then the result of that experiment does not provide evidence one way or another toward the truth of the hypothesis.*

The basic idea is that the data set should have a chance to falsify you. While this axiom makes intuitive sense, we can provide no justification for it. My claiming that “the sky is either blue or not blue” and then going out and verifying that the sky is indeed blue, hence that I was correct, sheds no light on my intelligence or ability to predict weather events.

**Example** (Postal Scam): Suppose that for 5 weeks in a row, a letter arrives in the mail that predicts the outcome of the upcoming Monday night football game. You keenly watch each Monday and to your surprise, the prediction was correct each time. On the day after the fifth game, a letter arrives, stating that if you wish to see next weeks prediction, a payment of \$50.00 is required. Should you pay? We can approach this problem through the axiom of non-falsifiability. Your goal is to deduce whether this individual is a good predictor. Suppose that the cost of printing and mailing out each letter would be \$0.50. Then the hypothesis is not falsifiable as follows. On week 1 he mails 16 people one result to that weeks game and 16 people the other. Watching the outcome of the game, his prediction will be correct for 16 people. To 8

of these 16 he mails one prediction of week 2's game and to the other 8, the other prediction. He continues in this way for 5 weeks, at which point there is one person to whom he has mailed the correct prediction each week. That happens to be you! Now, including the final letter asking for payment, he has mailed  $2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 + 1 = 64$  letters. Thus it costs him \$32. If he is getting back \$50, it is feasible for him to implement this hypothesis class, which is not falsifiable by the observed data set, and so whether or not you pay \$50 for the next prediction should not be influenced by the fact that he got 5 weeks correct in a row! A little counter intuitive, and thus the idea behind many a successful and lucrative postal scam. If on the other hand, he is only asking for \$10, say, then you know that he could not have implemented a non-falsifiable prediction algorithm and hence the data supports the fact that it may be worth paying.

We can apply the axiom of non-falsifiability to the understanding of generalization for our general learning scenario as follows. We would like to make statements of the form “a hypothesis exists in our learning model with a certain value of  $\pi$ ” and to do this we demonstrate the existence of a function with a training error of  $\nu \approx \pi$ , and hence conclude that that function has a test error of  $\approx \pi$  (i.e., we have good generalization). We can only make such conclusions if our hypothesis was falsifiable by the data. More specifically, on a given data set of size  $N$ , there are at most  $2^N$  possible classifications of the data. If every one of these classifications is implementable by a function in our learning model, then for every  $\nu$ , a function with that value of  $\nu$  must exist, hence, our hypothesis is not falsifiable and we cannot make any generalization conclusions. Thus, from the falsifiability point of view, we see that what matters is how many of the  $2^N$  classifications our learning model can implement, independent of the number of functions in the learning model. This is a big step. We have jumped from the infinite class of functions to studying how many of the classifications it can do, a finite number.

**Example** (Perceptron in 2-d): For the perceptron in  $d=2$ , one only needs to have 3 points to guarantee that the data set can falsify you (figure (a)).



However, it is very unlikely that the 3 points lie as shown in the figure, on a straight line. On the other hand, any set of 4 points can falsify this learning model. In fact this learning model can implement at most 14 of the possible classifications on 4 data points.

In order to guarantee that a data set could falsify your learning model, it is necessary that every possible data set be able to falsify it. By requiring that every possible data set be able to falsify our learning model, we are effectively removing the input distribution from the picture. We are thus led to the following definitions.

**Definition 5.2 (Dichotomy)** Given a data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  of size  $N$  a dichotomy is an assignment  $\delta = \{\delta_1, \dots, \delta_N\}$  to each data point, where  $\delta_i = \pm 1$ . There are  $2^N$  different dichotomies.

**Definition 5.3 (Dichotomize)** A learning model  $\mathcal{L}$  is said to dichotomize the dichotomy  $\delta$  on the data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  if there exists a  $g \in \mathcal{L}$  such that  $g(\mathbf{x}_i) = \delta_i$  for all  $i = 1, \dots, N$ . One also says that  $\mathcal{L}$  implements or separates the dichotomy  $\delta$

**Definition 5.4 (Shatter)** The learning model is said to shatter the set of points  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ , if it can dichotomize every one of the  $2^N$  dichotomies on the set of points.

**Definition 5.5** ( $\Delta_{\mathcal{L}}(\mathbf{x}_1, \dots, \mathbf{x}_N)$ )  $\Delta(\mathbf{x}_1, \dots, \mathbf{x}_N)$  is the number of dichotomies that the learning model  $\mathcal{L}$  can implement on the data set  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .  $\Delta(\mathbf{x}_1, \dots, \mathbf{x}_N) \leq 2^N$

If  $\Delta_{\mathcal{L}}(\mathbf{x}_1, \dots, \mathbf{x}_N) < 2^N$ , then this particular data set can falsify our learning model. If  $\Delta_{\mathcal{L}}(\mathbf{x}_1, \dots, \mathbf{x}_N) \ll 2^N$  then it is more likely that the learning model is falsified, and thus the conclusion that good generalization will occur for a particular hypothesis is sounder.

**Definition 5.6 (Growth Function,  $m_{\mathcal{L}}(N)$ )** The growth function is the maximum number of dichotomies that the learning model can implement on any set of  $N$  points.

$$m_{\mathcal{L}}(N) = \max_{\mathbf{x}_1, \dots, \mathbf{x}_N} \Delta_{\mathcal{L}}(\mathbf{x}_1, \dots, \mathbf{x}_N) \quad (8)$$

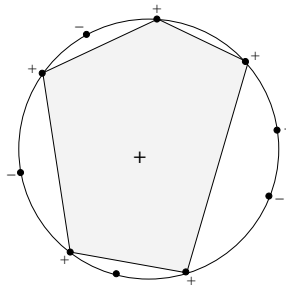
$$m_{\mathcal{L}}(N) \leq 2^N.$$

If  $m_{\mathcal{L}}(N) = 2^N$  then there exists a data set of size  $N$  that is shattered by  $\mathcal{L}$ . If  $m_{\mathcal{L}}(N) < 2^N$  then every data set of size  $N$  can falsify  $\mathcal{L}$ . The smaller  $m_{\mathcal{L}}(N)$ , the more likely that the data set can falsify you and the better the generalization should be. Without knowing the identity of the data set, we can bound the probability that the data set can falsify the learning model  $\mathcal{L}$  from below by (assuming all possible classifications of the data points are a priori equally likely)

$$P[\text{falsification}] \geq 1 - \frac{m_{\mathcal{L}}(N)}{2^N} \quad (9)$$

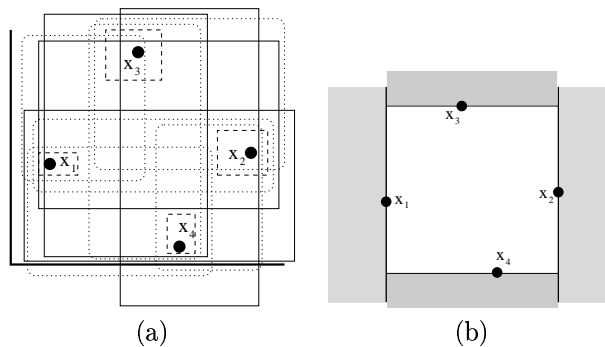
We see that if  $m_{\mathcal{L}} = o(2^N)$  then the right hand side approaches 1 and with very high probability, every data set could have falsified the learning model, so we can be confident that good generalization will occur. We now consider some examples.

1. **[Positive Ray]:** The learning model consists of functions of the form  $\text{sign}(x - a)$  in one dimension. Given  $N$  points, either the first 0, 1,  $\dots$ ,  $N$  points can be classified negative hence  $m(N) = N + 1$ .
2. **[Positive Interval]:** The learning model consists of functions that are positive on some interval and negative elsewhere. There are  $N + 1$  regions defined by the  $N$  points. If the end points of the positive interval are in the same region then all points are classified negative. The  $\binom{N+1}{2}$  ways of placing the endpoints of the positive interval in two different regions each yield a different dichotomy, hence  $m(N) = 1 + \binom{N+1}{2}$ .
3. **[Convex Sets]:** The learning model consists of functions that are positive inside some convex set in 2 dimensions and negative elsewhere. In this case,  $\Delta(\mathbf{x}_1, \dots, \mathbf{x}_N)$  can drastically depend on the location of the points. However, if no point is in the convex hull of the remaining points, then every dichotomy can be implemented. One possible arrangement is to arrange the points on a circle as shown in the figure.



Thus for every  $N$  there is an arrangement of the points for which all  $2^N$  dichotomies can be implemented, hence,  $m(N) = 2^N$ . Clearly,  $m(N)$  is an overestimate of the number of implementable dichotomies, for how likely is it that every point will not be contained in the convex hull of the remaining points. The specific arrangement of the points is a function of the input distribution and if we want to remain independent of the input distribution, then this is the price we have to pay.

4. **[Positive Rectangles]:** The learning model consists of functions that are positive on some rectangle in 2 dimensions and negative elsewhere. Consider  $m(4)$ . Figure (a) shows an arrangement of 4 points and the rectangle positions that would implement any dichotomy, thus  $m(4) = 16$ . What about  $m(5)$ ? In figure (b), it is clear that adding a 5th point would allow at least one dichotomy that could not be implemented. That there is no arrangement of 5 points for which all dichotomies can be implemented is left as an exercise for the reader.



Thus, though we have not computed  $m(5)$ , we know that  $m(5) < 2^5$ . In fact, it would be a daunting task to compute  $m(N)$  for general  $N$  for this apparently simple learning model.

The last example illustrates the fact that the computation of  $m(N)$  is quite a daunting task. In fact, generally computing  $m(N)$  is usually not possible, however, it might be useful to simply obtain a bound for  $m(N)$ . After all, all we need is that  $m(N) = o(2^N)$  to get good generalization in the limit of large  $N$  so any sub-exponential bound would work.

One observation about  $m(N)$  that can be made immediately is that if  $m(N^*) < 2^{N^*}$  for some  $N^*$ , then for all  $M \geq N^*$ ,  $m(M) < 2^M$ . This is because if  $m(M) = 2^M$  for some  $M > N^*$  then there is a set of data points that are shattered by the learning model, in which case any subset of these points (in particular, one of size  $N^*$ ) is shattered, hence  $m(N^*) = 2^{N^*}$ , contradicting the fact that  $m(N^*) < 2^{N^*}$ . Thus, for the positive rectangle model, we are at least guaranteed falsifiability for every  $N \geq 5$ . Even more can be said, however. To illustrate, consider 6 points. The dichotomies implementable on any set of six points are such that no subset of 5 points can be shattered. It turns out that this combinatorial restriction is a severe one, severe enough to restrict the number of dichotomies to being at most polynomial in  $N$ . This is a serious breakthrough, and is essentially the main theorem, and quite a surprising one at that. We will prove a theorem that states that there are only two kinds of learning models (good and bad ones). Good ones, ones that at some number of data points become falsifiable, have a growth function bounded by a polynomial in  $N$ . Thus, not only do they become falsifiable at some point, but they become falsifiable with probability 1 as  $N$  becomes large, under the assumption that all dichotomies are equally likely. Bad learning models never become falsifiable, and hence  $m(N) = 2^N$  for all  $N$ . It seems that the first  $N$  for which a learning model becomes falsifiable is important, so we are in the mood for the following definition.

**Definition 5.7 (VC Dimension)** *The VC dimension,  $d_{VC}$  of a learning model, if it exists, is the unique number for which  $m(N) = 2^N$  for  $N \leq d_{VC}$  and  $m(N) < 2^N$  for  $N > d_{VC}$ . If such a number does not exist, then we say that  $d_{VC} = \infty$ .*

Thus to compute  $d_{VC}$  for a learning model, it suffices to find an  $N^*$  for which  $m(N^*) = 2^{N^*}$  and  $m(N^* + 1) < 2^{N^* + 1}$ . Then,  $d_{VC}$  is this  $N^*$ . In order to bound  $d_{VC}$ , it suffices to find an  $N^*$  for which  $m(N^*) < 2^{N^*}$  in which case  $d_{VC} < N^*$ . For the 4 examples we considered above, it is not hard to compute  $d_{VC}$ .

1. **[Positive Ray]:**  $d_{VC} = 1$ .
2. **[Positive Interval]:**  $d_{VC} = 2$ .
3. **[Positive Convex Set]:**  $d_{VC} = \infty$ .
4. **[Positive Rectangle]:**  $d_{VC} = 4$ .

It might appear that  $d_{VC}$  is equal to the number of parameters. This is not generally true, though it appears so from the above examples. The learning model  $g(x) = \text{sign}(\sin(ax))$  has only one parameter,  $a$ . However, its VC dimension is  $\infty$ ! We leave this as an exercise for the reader.

We are now ready for our main theorem.

**Theorem 5.8 (m(N) Bound)** *If  $d_{VC} < \infty$  then  $m_{\mathcal{L}}(N) \leq N^{d_{VC}} + 1$ . If  $d_{VC} = \infty$  then  $m_{\mathcal{L}}(N) = 2^N$  for all  $N$ .*

PROOF: The second claim is true by definition. We prove the first claim. A finite  $d_{VC}$  implies that if  $N > d_{VC}$ , then every subset of size greater than  $d_{VC}$  cannot be shattered. Thus it seems convenient that we introduce the function  $B(N, d)$  defined as follows.

**Definition 5.9 ( $B(N, d)$ )** *Let  $B(N, d)$  be the maximum number of dichotomies on  $N$  points such that no subset of size  $d + 1$  is shattered.*

It is clear that  $m(N) \geq B(N, d)$  for all  $N$  and  $d$ . Further, if  $d \geq d_{VC}$  then  $m(N) \leq B(N, d)$  as the restriction requiring that no subset of size  $d + 1$  is shattered is a vacuous restriction given that  $d_{VC} \leq d$ . Hence we conclude that  $m(N) = B(N, d_{VC})$ , thus if we can analyse  $B(N, D)$  for all  $N, d$  then we can extract  $m(N)$ . To warm up, we consider some simple cases.  $B(N, 0) = 1$  for all  $N$  since if no subset of size 1 is to be shattered, then only one dichotomy can be allowed as a second different dichotomy must differ on at least one data point and that subset of size 1 is shattered.  $B(1, d) = 1$  for  $d = 0$  and  $B(1, d) \leq 2$  for  $d > 0$  (since  $B(N, d) \leq 2^N$ ). If there are more than 2 different functions in our learning model, then  $B(1, d) = 2$  for  $d > 0$ . We now move to the general case and try to develop a recursion for  $B(N, d)$  as follows. Suppose we have computed  $B(N + 1, d)$ . We can list the  $B(N + 1, d)$  different dichotomies that can be implemented for some set of points  $\mathbf{x}_1, \dots, \mathbf{x}_N$ .

		Dichotomies	$\mathbf{x}_1$	$\mathbf{x}_2$	...	$\mathbf{x}_{N-1}$	$\mathbf{x}_N$	$\mathbf{x}_{N+1}$
$S_1^+$	$\delta_1$	+1	-1	...	+1	+1	+1	+1
	$\delta_2$	-1	-1	...	+1	+1	+1	+1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\delta_{S_1-1}$	+1	-1	...	+1	-1	+1	+1
	$\delta_{S_1}$	-1	-1	...	-1	+1	+1	+1
$S_1^-$	$\delta_{S_1+1}$	+1	-1	...	+1	+1	-1	-1
	$\delta_{S_1+2}$	-1	-1	...	+1	+1	-1	-1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\delta_{2S_1-1}$	+1	-1	...	+1	-1	-1	-1
	$\delta_{2S_1}$	-1	-1	...	-1	+1	-1	-1
$S_2$	$\delta_{2S_1+1}$	+1	+1	...	+1	+1	+1	+1
	$\delta_{2S_1+2}$	+1	+1	...	+1	-1	-1	-1
	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$\delta_{B(N+1,d)-1}$	+1	-1	...	-1	-1	-1	-1
	$\delta_{B(N+1,d)}$	-1	-1	...	-1	-1	-1	+1

We have chosen a convenient order in which to list the dichotomies as follows. Consider the dichotomies that appear on the first  $N$  points. Some dichotomies on these  $N$  points appear twice, once with  $\mathbf{x}_{N+1} = +1$

and once with  $\mathbf{x}_{N+1} = -1$ . We collect these dichotomies in the set  $S_1$  which can be divided into two equal parts,  $S_1^+$  and  $S_1^-$ . The remaining dichotomies on the first  $N$  points that only appear once are collected in the set  $S_2$ . When it is unambiguous, we will use the names of the sets to refer to their sizes as well as to their contents. Clearly,  $B(N+1, d) = S_1^+ + S_1^- + S_2$ . The total number of dichotomies on the first  $N$  points is given by  $S_1^+ + S_2$ . Since no subset of  $d+1$  of these first  $N$  points can be shattered (since no  $d+1$  subset of all  $N$  points can be shattered), we deduce that  $S_1^+ + S_2 \leq B(N, d)$ . Further, no subset of size  $d$  of the first  $N$  points can be shattered by the dichotomies in  $S_1^-$ , as if there existed such a subset, then taking the corresponding set of dichotomies in  $S_1^+$  and adding  $\mathbf{x}_{N+1}$  to the data points yields a subset of size  $d+1$  that is shattered, which cannot happen by construction of  $B(N+1, d)$ . Hence,  $S_1^- \leq B(N, d-1)$ . We have thus proved that

$$B(N+1, d) \leq B(N, d) + B(N, d-1) \quad (10)$$

We can thus construct a bound for  $B(N, d)$  as shown in the following table.

	$d$						
	0	1	2	3	4	5	...
1	1	2	2	2	2	2	...
2	1	3	4	4	4	4	...
3	1	4	7	8	8	8	...
4	1	5	↓	11	...	...	...
5	1	6	⋮	⋱			
6	1	7	⋮		⋱		
⋮	⋮	⋮	⋮			⋱	

To summarize,  $m(N) = B(N, d_{VC})$  and

**Property 1:**  $B(N, 0) = 1$ ,  $B(1, d) \leq 2$  and  $B(N+1, d) \leq B(N, d) + B(N, d-1)$ .

The following two lemmas are useful.

**Lemma 5.10** *Any function satisfying property 1 above can be bounded as follows*

$$B(N, d) \leq \sum_{i=0}^d \binom{N}{i} \quad (11)$$

PROOF: The proof is by induction on  $N$ . Let the induction statement be

**P(M):** The bound on  $B(N, d)$  is valid for all  $N \leq M$  and all  $d$ .

It is clear that  $P(1)$  is true. Suppose that  $P(M)$  is true. and consider the statement  $P(M+1)$ . We need to consider  $B(M+1, d)$  for all  $d$ . If  $d = 0$  the bound is true. Consider  $d > 0$ . By property 1,  $B(M+1, d) \leq B(M, d) + B(M, d-1)$  and since  $P(M)$  is true, we find that

$$\begin{aligned} B(M+1, d) &\leq \sum_{i=0}^d \binom{M}{i} + \sum_{i=0}^{d-1} \binom{N}{i} = 1 + \sum_{i=1}^d \binom{N}{i} + \sum_{i=1}^d \binom{N}{i-1} \\ &= 1 + \sum_{i=1}^d \left( \binom{N}{i} + \binom{N}{i-1} \right) \\ &= 1 + \sum_{i=1}^d \binom{N+1}{i} = \sum_{i=0}^d \binom{N+1}{i} \end{aligned}$$

Where the identity  $\binom{N+1}{i} = \binom{N}{i} + \binom{N}{i-1}$  has been used. This identity can be proven by noticing that to obtain the number of ways to pick  $i$  objects from  $N+1$ , either the first object is included (in  $\binom{N}{i-1}$  ways)

or the first object is not included (in  $\binom{N}{i}$  ways). Thus  $P(M+1)$  is true and so by induction  $P(M)$  is true for all  $M$ .  $\square$

**Lemma 5.11**

$$\sum_{i=0}^d \binom{N}{i} \leq N^d + 1$$

PROOF: Again, the proof is by induction, this time on  $d$ . Let the induction statement be

**P(D):** The bound on  $\sum_{i=0}^d \binom{N}{i}$  is valid for all  $d \leq D$  and all  $N$ .

When  $D = 0$ , the bound yields 2 and the sum gives 1. Thus  $P(0)$  is true. Suppose that  $P(D)$  is true. Now consider  $P(D+1)$ . We need to consider  $\sum_{i=0}^{D+1} \binom{N}{i}$  for all  $N$ . But, by the induction hypothesis,

$$\sum_{i=0}^{D+1} \binom{N}{i} = \sum_{i=0}^D \binom{N}{i} + \binom{N}{D+1} \leq 1 + N^D + \frac{N!}{(N-D-1)!(D+1)!} \quad (12)$$

The bound is valid for  $N = 1$ . For  $1 < N \leq D$ , the last term is zero and the bound will be valid. For  $N > D$ ,  $\binom{N}{D+1} \leq N^{D+1}(1-1/N)$  which upon plugging into (12) yields  $1 + N^{D+1}$ , thus the bound is valid for all  $N \geq 1$  and so  $P(D+1)$  is true, concluding the proof.  $\square$

The theorem now follows by a straight forward application of these two lemmas.  $B(N, d) \leq N^d + 1$  for all  $N, d$ . in particular  $m(N) = B(N, d_{VC}) \leq N^{d_{VC}} + 1$ .  $\blacksquare$

We have spent some time discussing  $m(N)$  and its significance through the axiom of falsifiability. Based upon this discussion, it appears that learning models come in two forms, good ones and bad ones. Good ones have  $d_{VC} < \infty$ . These learning models are falsifiable with probability 1 in the large  $N$  limit, since  $m(N)$  grows at most polynomially in  $N$ . We expect good generalization to occur for these learning models when  $N$  becomes large (from the point of view of the axiom of non-falsifiability). It is easy to check that the bound on  $m(N)$  applies for the 4 examples we have been considering. Bad learning models are never falsifiable and we never expect good generalization to occur.

It is now time to put some meat onto these statements. In other words, is  $m(N)$  the only quantity we need to look at in order to guarantee good generalization. Is our axiom of non-falsifiability really faithful, i.e., is falsifiability alone enough to guarantee good generalization? We go back to the generalization question. We would like to obtain the right hand side for (7). Luckily, someone else did the hard work, but the result should not be surprising by now. It turns out that falsifiability is enough. Falsifiable learning models give good generalization. At this point the following theorem should not be too surprising.

**Theorem 5.12 (Vapnik-Chervonenkis, Parrondo-Van den Broek)**

$$P \left[ \sup_{g \in \mathcal{L}} |\nu_g - \pi_g| \geq \epsilon \right] \leq 6e^{2\epsilon} m_{\mathcal{L}}(2N) e^{-\epsilon^2 N} \quad (13)$$

PROOF: See [1], [5].  $\blacksquare$

Notice the similarity between this bound and the one for the finite  $M$  bound (5). Due to minor technicalities that arise in the proof, the factor of 2 becomes  $6e^{2\epsilon}$  and we identify  $M_{eff}$  as  $m_{\mathcal{L}}(2N)$ . From our earlier arguments, we might have expected  $m_{\mathcal{L}}(N)$ , but once again, due to the technicalities in the proof,  $m_{\mathcal{L}}(2N)$  appears.

Lets take a moment to examine the slack in this bound. From the technical point of view, the slack arises in the following ways:

1. Use of Chernoff Bound.
2. Use of Union bound (in multiplying by  $m(N)$ ).
3. Use of polynomial bound for  $m(N)$

4. Insistence on distribution independence by using  $m(N)$  rather than for example  $\Delta(\mathbf{x}_1, \dots, \mathbf{x}_N)$  or its expected value.

Some effort could be put into tightening the bound or invoking some input distribution dependence, but the eventual reality is that this bound is extremely loose and should be regarded as providing for the *possibility* of learning in the large  $N$  limit rather than as an actual quantitative bound given  $N$ . However, it appears that though the bound is loose, it is equally loose for different learning models, and hence is useful for comparing the generalization performance of different learning models. Thus, applying the bound to  $\mathcal{L}_1$  and  $\mathcal{L}_2$  might yield bounds greater than 1 in both cases, but if the bound for  $\mathcal{L}_1$  is significantly less than that for  $\mathcal{L}_2$ , one can usually expect better generalization performance from  $\mathcal{L}_1$ . It should be noted that none of these statements will stand up to rigorous scrutiny. That having been said, . . .

## 5.2 Using the VC Bound

For learning models with finite  $d_{VC}$ , we can replace  $m_{\mathcal{L}}(2N)$  with the polynomial bound for it, obtaining

$$P \left[ \sup_{g \in \mathcal{L}} |\nu_g - \pi_g| \geq \epsilon \right] \leq 6e^{2\epsilon} ((2N)^{d_{VC}} + 1) e^{-\epsilon^2 N} \quad (14)$$

The two approaches to using this inequality arise from the following two ways of interpreting it:

1. **Sample Complexity:** *The probability of a large deviation is small (can be bounded from above) for sufficiently large  $N$ .* In other words, fix  $(\epsilon, \eta)$ . We can now ask, how large must  $N$  be to guarantee that the probability of a generalization error greater than  $\epsilon$  is less than  $\eta$ . This is called the sample complexity of the learning model  $N(\epsilon, \delta)$ . If  $d_{VC} < \infty$ , we can bound  $N(\epsilon, \eta)$  by requiring that

$$6e^{2\epsilon} m_{\mathcal{L}}(2N) e^{-\epsilon^2 N} \leq \eta \quad (15)$$

which upon solving for  $N$  yields that

$$N \geq \frac{1}{\epsilon^2} \log \left( \frac{6e^{2\epsilon} m_{\mathcal{L}}(2N)}{\eta} \right) \quad (16)$$

Replacing  $m_{\mathcal{L}}(2N)$  by its upper bound in the above equation would also yield a valid bound on the sample complexity. It is now necessary to solve this inequality for  $N$  and a number of iterative methods exist.

2. **Test Error Bound:** *The probability of a small (at most  $\epsilon$ ) deviation is large (can be bounded from below) for sufficiently large  $\epsilon$ .* In other words, fix  $(N, \eta)$  and determine the smallest  $\epsilon$  such that the probability that the generalization error is less than  $\epsilon$  is greater than  $1 - \eta$ . Re-writing (14) we see that

$$P \left[ \sup_{g \in \mathcal{L}} |\nu_g - \pi_g| < \epsilon \right] \geq 1 - 6e^{2\epsilon} m_{\mathcal{L}}(2N) e^{-\epsilon^2 N} \quad (17)$$

Thus we need that

$$\eta \geq 6e^{2\epsilon} m_{\mathcal{L}}(2N) e^{-\epsilon^2 N} \quad (18)$$

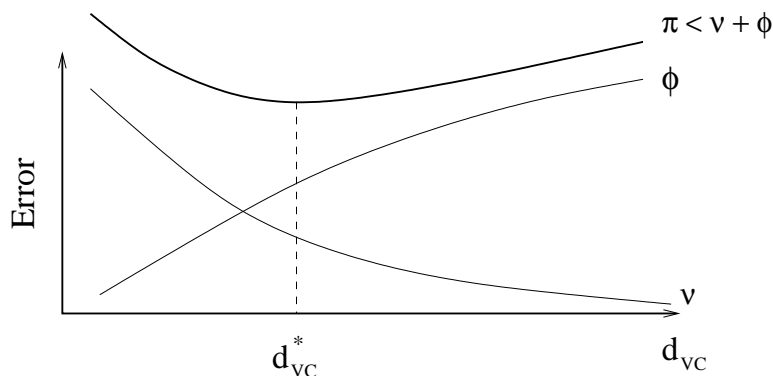
which upon solving for  $\epsilon$  (ignoring the  $e^{2\epsilon}$  term as it is approximately 1) yields

$$P \left[ \sup_{g \in \mathcal{L}} |\nu_g - \pi_g| < \epsilon \right] \geq 1 - \eta \quad \Rightarrow \quad \epsilon \geq \sqrt{\frac{1}{N} \log \left( \frac{6m_{\mathcal{L}}(2N)}{\eta} \right)} \quad (19)$$

thus using the smallest allowed  $\epsilon$  we see that with probability greater than  $1 - \eta$ ,

$$\nu - \sqrt{\frac{1}{N} \log \left( \frac{6m_{\mathcal{L}}(2N)}{\eta} \right)} \leq \pi \leq \nu + \sqrt{\frac{1}{N} \log \left( \frac{6m_{\mathcal{L}}(2N)}{\eta} \right)} \quad (20)$$

Usually it is only the upper bound that is of interest. Further replacing  $m_{\mathcal{L}}(2N)$  by its bound also yields a valid test error bound. Thus, we notice that the bound on the test error is composed of two parts. The training error plus a complexity term that increases as  $d_{VC}$  increases. The minimum achievable training error is expected to decrease as one increases  $d_{VC}$  hence the bound for the test error should have the behavior as shown in the next figure



One approach to picking the “optimal” number of hidden units is to pick the value of  $d_{VC}$  that minimizes this bound, in the hopes that the minimum of this bound is attained at about the minimum of the test error itself. Notice that once the complexity of the learning model is fixed (at  $d_{VC}$ ), the complexity term of the bound is fixed. Thus the optimal next step is to actually obtain the *minimum*  $v$ !

## References

- [1] J. M. R. Parrondo and C. Van den Broeck. Vapnik-chervonenkis bounds for generalization. *Journal of Physics A: Math. Gen.*, 26:2211–2223, 1993.
- [2] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer Series in Statistics. Springer Verlag, New york, 1982.
- [3] V. N. Vapnik. *The Nature of Statistical Learning Theory*. Springer–Verlag, 1995.
- [4] V. N. Vapnik. *Statistical Learning Theory*. Adaptive and Learning Systems for Signal Processing, Communications and Control. John Wiley & Sons, Inc., New york, 1998.
- [5] V. N. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16:264–280, 1971.