Machine Learning - I (CS/DS 864)
Aug–Dec
***Instructor*: Prof. G.Srinivasaraghavan**

Lecture Notes

Generalization Theory

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Contents

List of Figures

List of Tables

In this document we examine one of the two central question in learning that were highlighted in the document on the Learning Problem formulation — "When can we expect to see a predictable generalization of a hypothesis to test samples outside the training dataset?" In other words we ask "When is PAC learning feasible?" We present the definitive answer to this question in the ERM setting. These results are predominantly due to Vladimir Vapnik and Alexei Chervonenkis from the path-breaking work they carried out in the 1970s.

1 PAC Learning of Finite Hypothesis Spaces

We illustrate the idea by showing that any finite hypothesis space is PAC learnable. In particular we show that for any finite hypothesis class $\mathcal{H}$, the probability of some hypothesis $h \in \mathcal{H}$ not generalizing is small — with a large enough sample the probability (with respect to the training sample) of not being able to produce a hypothesis whose empirical risk on the training sample is more than $0 < \epsilon < 1$ away from the true risk of the hypothesis can be made arbitrarily small. Formally

Theorem 1 *For any learning algorithm $\mathcal{A}$ with the hypothesis class $\mathcal{H}$ such that $|\mathcal{H}| = k$ for a fixed constant k, if $h^* = \mathcal{A}(\mathcal{D}_n)$ is the output hypothesis on a training dataset $\mathcal{D}_n$ of size n then*

$$Pr(|R_e(h^*) - R(h^*)| > \epsilon) \leq 2k.e^{-2\epsilon^2 n}$$

Proof: We will bound the probability that the hypothesis h^* will not generalize — i.e., the probability that the true risk $R(h^*)$ of h^* across the domain $\mathcal{D}$ is too far away from the empirical risk $R_e(h^*)$ measured on $\mathcal{D}_n$.

$$\begin{aligned} Pr(|R_e(h^*) - R(h^*)| > \epsilon) &\leq Pr\left(\bigvee_{i=1}^{k}(|R_e(h_i) - R(h_i)| > \epsilon)\right) \\ &\leq \sum_{i=1}^{k} Pr(|R_e(h_i) - R(h_i)| > \epsilon) \\ &\leq k.\sup_{h\in\mathcal{H}} Pr(|R_e(h) - R(h)| > \epsilon) \\ &= k.Pr(|R_e(\hat{h}) - R(\hat{h})| > \epsilon) \end{aligned}$$

where $\hat{h}$ is the 'worst case' hypothesis — one for which the probability of not generalizing is the highest. Note that the probability in the inequalities above is over all possible datasets $\mathcal{D}_n$. The worst case hypothesis $\hat{h}$ is clearly independent of the specific training dataset chosen.

Consider a fixed hypothesis $\hat{h}(\boldsymbol{x}) : \mathcal{D} \to \pm 1$ that attempts to approximate an unknown target function $f(\boldsymbol{x}) : \mathcal{D} \to \pm 1$. The points $\boldsymbol{x}$ are distributed according to an unknown distribution $p_{\mathcal{D}}(\boldsymbol{x})$. Let X be a Bernoulli random variable that maps every point $\boldsymbol{x}$ to a value $\{0, 1\}$ as follows:

$$X(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \hat{h}(\boldsymbol{x}) = f(\boldsymbol{x}) \\ 1 & \text{Otherwise} \end{cases}$$

It is easy to verify that

$$E[X] = R(\hat{h}) \quad \text{and} \quad \frac{1}{n}\sum_{\boldsymbol{x}\in\mathcal{D}_n} X(\boldsymbol{x}) = R_e(\hat{h})$$

We can now apply well known tail bounds (Hoeffding's inequality [**hoeffding**] in particular) to bound the difference between the average over an empirical sample of size n for the random variable X and the true expected value $E[X]$. Denoting the sample of n values for the random variable as $X_1, \ldots, X_n$, Hoeffding's inequality for a Bernoulli random variable can be written as

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - E[X]\right| > \epsilon\right) \leq 2.e^{-2\epsilon^2 n}$$

Applying this to $\hat{h}$ we get

$$Pr(|R_e(\hat{h}) - R(\hat{h})| > \epsilon) \leq 2.e^{-2\epsilon^2 n}$$

The following bound now follows:

$$Pr(|R_e(h^*) - R(h^*)| > \epsilon) \leq 2k.e^{-2\epsilon^2 n}$$

■

The bound of the theorem above can actually be improved significantly by taking k as just the number of equivalence classes of hypotheses in $\mathcal{H}$ — each equivalence class consists of all the hypotheses that induce identical ± 1 labellings on the dataset $\mathcal{D}_n$. With this the bound now can potentially extend even to infinite hypothesis classes with k replaced by the number of equivalence classes. The number of distinct labellings that can be generated on a set of n data points by a the hypotheses in a hypothesis class is called the *Growth Function* of that class. We examine the the notion of a growth function in detail in the next section and use that to explore the fundamental theorem of learning which gives a necessary and sufficient condition for the feasibility of PAC learning.

2 Growth Functions and the VC-Dimension

Consider a set of data points $S = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ in $\mathcal{R}^d$. Let $\mathcal{F}$ be a class of labelling functions $f(\mathbf{x}) : \mathcal{R}^d \to \pm 1$. S is said to be *shattered* by $\mathcal{F}$ if for any assignment $\boldsymbol{x}_i \to y_i \in \pm 1$ of ± 1 labels to the points in S there exists some $g \in \mathcal{F}$ such that $\forall i : g(\boldsymbol{x}_i) = y_i$. Note that there can be 2^n distinct labellings of S. The accepted term that is used for a labelling with 2 labels is a *dichotomy*. Therefore $\mathcal{F}$ is said to shatter S if the number of distinct dichotomies of S produced by the functions in $\mathcal{F}$ is exactly 2^n.

Definition 1 ***VC-Dimension****: VC-Dimension (Vapnik-Chervonenkis Dimension)* $\mathsf{d}_{\mathcal{F}}^{vc}$ *of a class of functions* $\mathcal{F}$ *is the largest* n *such that* $\mathcal{F}$ *shatters some set of* n *points. In other words* $\mathsf{d}_{\mathcal{F}}^{vc}$ *is largest number of points that can be arranged in such a way that* $\mathcal{F}$ *shatters them.*

In the rest of this document we will omit the superscript *vc* from the notation for the VC-dimension of a class of functions $\mathcal{F}$ and simply denote it as $\mathsf{d}_{\mathcal{F}}$.

Definition 2 ***Growth Function****: Growth Function $\mathcal{G}_{\mathcal{F}}(n)$ of a class of functions $\mathcal{F}$ is the maximum number of labellings of n points that can be generated by the functions in $\mathcal{F}$. It should be easy to see from the definition of $\mathsf{d}_{\mathcal{F}}$ that*

$$\mathcal{G}_{\mathcal{F}}(n) \begin{cases} = 2^n & \text{if } n \leq \mathsf{d}_{\mathcal{F}} \\ < 2^n & \text{otherwise} \end{cases}$$

2.1 Examples of VC Dimension

We introduce the geometric notion of a simplex. We will refer to this in some of the examples in this section. A *Simplex* in $\mathcal{R}^d$ is the simplest possible polytope in $\mathcal{R}^d$. It can be inductively defined as follows:

- A line-segment is a 1-d simplex and a triangle is a 2-d simplex. The 3-d simplex is the tetrahedron. See Figure 1.
- Form a $(d+1)$-dimensional simplex by taking a d-simplex Δ^d in $\mathcal{R}^{(d+1)}$ and a point $\boldsymbol{p}$ outside the hyperplane in $\mathcal{R}^{(d+1)}$ containing Δ^d and joining every point of Δ^d with $\boldsymbol{p}$.

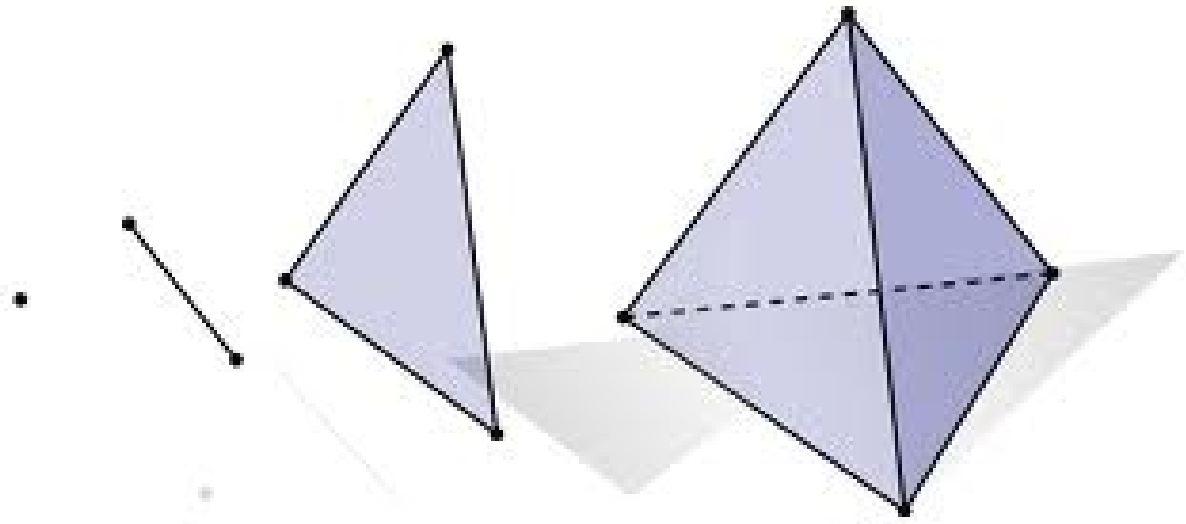

Figure 1: Simplices in 0, 1, 2 and 3 Dimensions

Exercise 1 *Show the following interesting properties of simplices in $\mathcal{R}^d$.*

1. *The number of vertices of a simplex is exactly $(d+1)$.*

2. *It has exactly* $(d+1)$ *facets (faces of dimension* $(d-1)$*) each of which is a* $(d-1)$*-simplex.*

3. *Every subset with* k *of its vertices forms a* $(k-1)$*-simplex.*

Exercise 2 *Show that the following alternate characterization of a simplex is equivalent to the construction we gave above. Any* $(d+1)$ *points* $\{\boldsymbol{v}_0, \boldsymbol{v}_1, \ldots, \boldsymbol{v}_d\}$ *in* $\mathcal{R}^d$ *form the corners of a simplex if and only if the vectors*

$$(\boldsymbol{v}_1 - \boldsymbol{v}_0), (\boldsymbol{v}_2 - \boldsymbol{v}_0), \ldots, (\boldsymbol{v}_d - \boldsymbol{v}_0)$$

are linearly independent.

A few examples of function classes with the growth function and the VC-dimension for each are given below. In each case $\mathcal{F}$ refers to the function class with functions of the kind $f : \mathbb{R}^d \to \pm 1$.

1. **Hyperplanes in** $\mathcal{R}^d$: It is easier to examine this case in homogeneous coordinates. We therefore assume the points are from $\mathbb{R}^d$. Consider any set of $(d+2)$ points in $\mathbb{R}^d$. There must be one point $\boldsymbol{x}$ among the $(d+2)$ points that is a linear combination of the others. Hence $\boldsymbol{a}^T.\boldsymbol{x}$ is completely determined by the other points. The label $sign(\boldsymbol{a}^T.\boldsymbol{x})$ therefore cannot independently take both the values ± 1. Therefore $\mathsf{d}_{\mathcal{F}} < (d+2)$. To show that $\mathsf{d}_{\mathcal{F}} \geq d+1$ we only need to demonstrate a set of $(d+1)$ points that can be shattered by $\mathcal{F}$. Refer Figure 2.1 for some visual illustrations. In this case

$$\mathcal{F} = \{f \mid f(\boldsymbol{x}) = sign(\boldsymbol{a}^T.\boldsymbol{x}) \text{ for some } \boldsymbol{a} \in \mathcal{R}^{(d+1)}\}$$
$$\mathsf{d}_{\mathcal{F}} = (d+1)$$

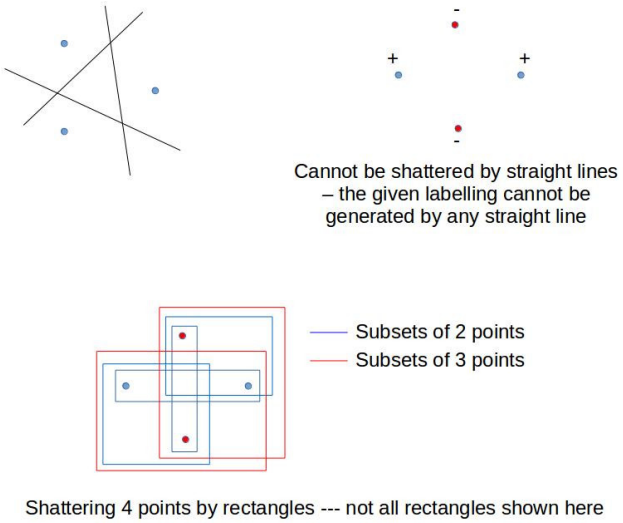

Figure 2: Shattering Examples

Exercise 3 *Show that the set of vertices of a simplex can be shattered by the class of hyperplane separators.* ***Hint****: Use induction on* d*.*

Exercise 4 *Show that when* $d = 1$ *the growth function for this class of functions is* $2n$ *where* n *is the size of the data set.*

2. **Simplices in** $\mathcal{R}^d$: In this case

$$\mathcal{F} = \left\{f \mid f(\boldsymbol{x}) = +1 \text{ iff } \boldsymbol{x} \in \Delta^d \text{ for some } d\text{-simplex } \Delta^d \in \mathcal{R}^d\right\}$$
$$\mathsf{d}_{\mathcal{F}} = d(d+1)$$

Exercise 5 *Show that when $d = 1$ the growth function for this class of functions is $\binom{n+1}{2} + 1$ where n is the size of the data set.*

3. **Axis-Parallel Rectangular Boxes in $\mathcal{R}^d$**: An axis-parallel rectangular box in $\mathcal{R}^d$ is a region consisting of all points bounded by a minimum and maximum in each coordinate direction. Formally the rectangular box defined by a set of $2d$ reals $l_i < h_i, i = 1, \ldots, d$ is the set

$$B(\boldsymbol{l}, \boldsymbol{h}) = \{\boldsymbol{x} \mid \bigwedge_{i=1}^{d} (l_i \leq x_i \leq h_i)\}$$

In this case

$$\mathcal{F} = \left\{f \mid f(\boldsymbol{x}) = +1 \text{ iff } \boldsymbol{x} \in B(\boldsymbol{l}, \boldsymbol{h}) \text{ for some } \boldsymbol{l}, \boldsymbol{h} \in \mathcal{R}^d\right\}$$
$$\mathsf{d}_{\mathcal{F}} = 2d$$

It is easy to show that $\mathsf{d}_{\mathcal{F}} \leq 2d$. Take any set S of $(2d+1)$ points. Let $\boldsymbol{y}_i, \boldsymbol{z}_i$ be the points such that

$$\forall i \leq d : y_{ii} = \min\{x_i \mid \boldsymbol{x} \in S\}, z_{ii} = \max\{x_i \mid \boldsymbol{x} \in S\}$$

There are at most $2d$ such distinct points. Clearly it is not possible to construct a rectangle that contains only the points $\{\boldsymbol{y}_i, \boldsymbol{z}_i, i = 1, \ldots, d\}$ and none of the others — in fact such a rectangle will contain all the points of S. Refer Figure 2.1 for some visual illustrations.

The vertices of a rhombus whose diagonals are parallel to the two axes is a dataset that can be shattered by a family of rectangles in 2-dimensions. See Figure 2.1.

Exercise 6 *Generalize the example above (rhombus) to construct a family of sets of $2d$ points in $\mathcal{R}^d$ for any $d > 1$, that can be shattered by axis parallel rectangular boxes.*

4. **Convex Sets in $\mathcal{R}^d$**: This can be established using the following construction. For any n choose a set S of n distinct points on a sphere. Consider any dichotomy over S — we can construct a polytope P (a convex set) whose vertices are the points of S with label +1. Clearly P will be entirely contained within the sphere. Hence the only points of S in P are the ones in S^+ by construction. S^- lies entirely outside P. See Figure 3 for an illustration. Therefore P will discriminate between S^+ and S^-. Hence any dichotomy of S can be generated by the set of all convex sets in $\mathcal{R}^d$. So in this case $\mathcal{G}_{\mathcal{F}}(n) = 2^n$ for every n where

$$\mathcal{F} = \left\{f \mid f(\boldsymbol{x}) = +1 \text{ iff } \boldsymbol{x} \in \mathcal{C} \text{ for some convex set } \mathcal{C} \subset \mathcal{R}^d\right\}$$
$$\mathsf{d}_{\mathcal{F}} = \infty$$

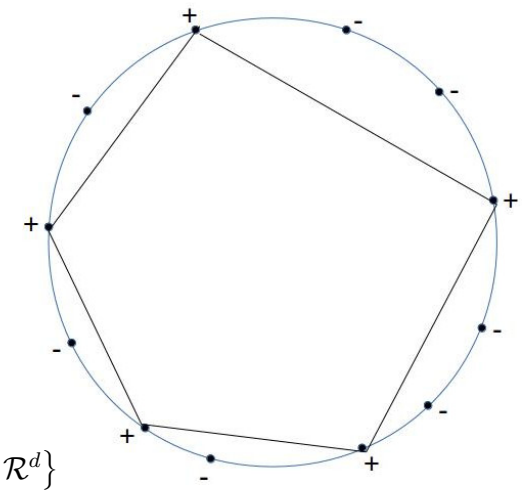

Figure 3: $\mathsf{d}_{\mathcal{F}}$ of a Convex Set

5. **Spherical Balls in $\mathbf{R}^d$**: A spherical ball $\mathcal{S}(\boldsymbol{c}, r)$ with center $\boldsymbol{c}$ and radius r is the set

$$\mathcal{S}(\boldsymbol{c}, r) = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{c}\| \leq r\}$$

In this case

$$\mathcal{F} = \left\{f \mid f(\boldsymbol{x}) = +1 \text{ iff } \boldsymbol{x} \in \mathcal{S}(\boldsymbol{c}, r) \text{ for some } \mathbf{c} \in \mathcal{R}^d, r > 0 \in \mathcal{R}\right\}$$
$$\mathsf{d}_{\mathcal{F}} = d + 1$$

Exercise 7 *Show that a d-simplex can be shattered by this class of functions.*

Exercise 8 *Show that one needs $(d+1)$ points to define a sphere in $\mathcal{R}^d$. Use this to show that no set of $(d+2)$ points in $\mathcal{R}^d$ can be shattered by a set of spherical balls.*

Exercise 9 *All the examples above except the one with convex sets, involve function classes that can be parametrized by a finite number of parameters — number of independent parameters that need to be specified to specify a particular function in the class. Verify that in each case $\mathsf{d}_{\mathcal{F}}$ is equal to the number of independent (free) parameters of the class.*

Exercise 10 *Construct a function class $\mathcal{F}$ parametrized by a single parameter, for which $\mathsf{d}_{\mathcal{F}} = \infty$.*

Exercise 11 *Show that the VC-dimension $\mathsf{d}_{\mathcal{F}}$ of a finite class of functions $\mathcal{F}$ is at most $\log_2 |\mathcal{F}|$.*

We conclude this section with a general result on the VC-dimension of a transformation.

Theorem 2 *Let $\mathcal{H}$ be a finite dimensional ($d_{\mathcal{H}}$) vector space of functions $h : \mathcal{R}^d \to \mathcal{R}$. Let $\mathcal{F}$ be the class of functions $f : \mathcal{R} \to \pm 1$ defined as*

$$\mathcal{F} = \{f \mid f(\mathbf{x}) = sign(h(\mathbf{x})),\ h \in \mathcal{H}\}$$

Then $\mathsf{d}_{\mathcal{F}} \leq d_{\mathcal{H}}$.

Proof: We show that no set of $(d_{\mathcal{H}}+1)$ points can be shattered by $\mathcal{F}$. Consider an arbitrary set of points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{(d_{\mathcal{H}}+1)} \in \mathcal{R}^d$. We show a dichotomy over any such set of points that cannot be generated by the functions in $\mathcal{F}$. Consider a linear transformation $\Gamma : \mathcal{H} \to \mathcal{R}^{(d_{\mathcal{H}}+1)}$ defined as

$$\Gamma(h) = (h(\mathbf{x}_1), h(\mathbf{x}_2), \ldots, h(\mathbf{x}_{(d_{\mathcal{H}}+1)}))$$

Dimensionality of $\mathcal{H}$ is $d_{\mathcal{H}}$ — hence after a linear transformation the dimensionality will remain at most $d_{\mathcal{H}}$. Therefore there exists a non-zero vector $\lambda \in \mathbf{R}^{(d_{\mathcal{H}}+1)}$ that is orthogonal to $\Gamma(\mathcal{H})$ — λ is orthogonal to every vector in $\Gamma(\mathcal{H})$. So

$$\forall h \in \mathcal{H} : \left(\sum_{i=1}^{(d_{\mathcal{H}}+1)} \lambda_i . h(\mathbf{x}_i)\right) = 0 \tag{1}$$

Label the points $\mathbf{x}_i$ as follows: $\mathbf{x}_i$ is labelled +1 if $\lambda_i > 0$ and −1 if $\lambda_i < 0$. We claim that this dichotomy cannot be realized by any function $f \in \mathcal{F}$. Suppose there was such a function, then there exists a $h^* \in \mathcal{H}$ such that $sign(h^*(\mathbf{x}_i)) = +1$ whenever $\mathbf{x}_i$ is labelled +1 and $sign(h^*(\mathbf{x}_i)) = -1$ whenever $\mathbf{x}_i$ is labelled −1. From our labelling scheme therefore each term of the summation in Equation 1 is ≥ 0 and not all terms are 0. Hence

$$\sum_{i=1}^{(d_{\mathcal{H}}+1)} \lambda_i . h^*(\mathbf{x}_i) > 0$$

violating Equation 1. ■

Exercise 12 *Show that for any two hypothesis classes $\mathcal{H}' \subset \mathcal{H}$, $\mathsf{d}_{\mathcal{H}'} \leq \mathsf{d}_{\mathcal{H}}$.*

Exercise 13 *Given set of labelled data points $\{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ where $y_i \in \pm 1$ from a distribution for which there exists a rectangular box $\boldsymbol{B}$ that contains all the data points with $y_i = +1$ in its interior. Design a linear time algorithm (linear in n) that determines a correct rectangular separator for the given set of points.*

Exercise 14 *Consider the restriction of the binary classification problem with axis parallel rectangles, in 1-dimension — every hypothesis is an interval on the real line, where the data points within the interval will be classified as +1. Come up with a dynamic programming algorithm to arrive at an interval that will minimize the number of misclassifications on a dataset with n values on the real line, each with a ± 1 label.*

Exercise 15 *Let $\mathcal{P}$ be the class of binary classifiers such that each classifier $f \in \mathcal{P}$ is of the form $f(x) = sign(p(x))$ for some univariate polynomial $p(x)$. Show that $\mathsf{d}_{\mathcal{P}} = \infty$.*

2.2 Relationship between Growth Functions and VC-Dimension

Lemma 1 ***Sauer's Lemma****: For a function class $\mathcal{F}$ with VC-dimension $\mathsf{d}_{\mathcal{F}}$*

$$\mathcal{G}_{\mathcal{F}}(n) \leq \sum_{i=0}^{\mathsf{d}_{\mathcal{F}}} \binom{n}{i}$$

Also if $n > \mathsf{d}_{\mathcal{F}}$

$$\mathcal{G}_{\mathcal{F}}(n) \leq \left(\frac{en}{\mathsf{d}_{\mathcal{F}}}\right)^{\mathsf{d}_{\mathcal{F}}} \Rightarrow \ln \mathcal{G}_{\mathcal{F}}(n) \leq \mathsf{d}_{\mathcal{F}} \left(\ln\left(\frac{n}{\mathsf{d}_{\mathcal{F}}}\right) + 1\right)$$

Note*: Sometimes $\ln \mathcal{G}_{\mathcal{F}}(n)$ is referred to as the Growth Function.*

Proof: Let the function $\mathbb{B}(n, k)$ represent the largest number of dichotomies that can be generated on a set S of n points in such a way that there is no subset of size k that is shattered. Since $\mathcal{F}$ cannot shatter any set of $(\mathsf{d}_{\mathcal{F}} + 1)$ points, $\mathbb{B}(n, \mathsf{d}_{\mathcal{F}} + 1)$ would serve as an upper bound on $\mathcal{G}(n, \mathsf{d}_{\mathcal{F}})$. We therefore prove the first part of the theorem by showing

$$\forall n, k : \mathbb{B}(n, k+1) = \sum_{i=0}^{k} \binom{n}{i} \Rightarrow \mathcal{G}_{\mathcal{F}}(n) \leq \mathbb{B}(n, \mathsf{d}_{\mathcal{F}} + 1) = \sum_{i=0}^{\mathsf{d}_{\mathcal{F}}} \binom{n}{i}$$

$\mathbb{B}(n, k+1) \geq \sum_{i=0}^{k} \binom{n}{i}$: We demonstrate a set of $\sum_{i=0}^{k} \binom{n}{i}$ dichotomies on n points without shattering any $(k+1)$-subset of the n points. That would imply the lower bound in the claim. Let $\mathcal{D}$ be the set of all dichotomies on n points with at most k points labelled +1. That $|\mathcal{D}| = \sum_{i=0}^{k} \binom{n}{i}$ is clear from the observation that $\binom{n}{i}$ is the number of subsets of size i. Also no subset of size $(k+1)$ is shattered because no subset of size $(k+1)$ can be labelled with all +1s.

$\mathbb{B}(n, k+1) \leq \sum_{i=0}^{k} \binom{n}{i}$: Let $\mathcal{D}_n$ be the set of $\mathbb{B}(n, k+1)$ dichotomies on a set S of n points. Let $\mathbf{x}_0$ denote one of the points in S and let $S_{(n-1)} = S_n - \{\mathbf{x}_0\}$. Let $\mathcal{D}_{(n-1)}$ be the projection of $\mathcal{D}_n$ on $S_{(n-1)}$ — $\mathcal{D}_{(n-1)}$ would simply be $\mathcal{D}_n$ with $\mathbf{x}_0$ removed. It is easy to see that $\mathcal{D}_{(n-1)}$ consists of 2 kinds of dichotomies $\mathcal{D}_{(n-1)} = \mathcal{D}^1_{(n-1)} \cup \mathcal{D}^2_{(n-1)}$ where

1. $\mathcal{D}^1_{(n-1)}$ is the set of dichotomies D on $S_{(n-1)}$ such that $D + (\mathbf{x}_0 \to +1) \in \mathcal{D}_n$ or $L + (\mathbf{x}_0 \to -1) \in \mathcal{D}_n$ but not both.
2. $\mathcal{D}^2_{(n-1)} = \{D \in \mathcal{D}_{(n-1)} \mid D + (\mathbf{x}_0 \to \pm 1) \in \mathcal{D}_n\}$.

Clearly $|\mathcal{D}_n| = |\mathcal{D}^1_{(n-1)}| + 2.|\mathcal{D}^2_{(n-1)}|$. Also the number of distinct dichotomies on $S_{(n-1)}$ induced by $\mathcal{D}_n$ is $(|\mathcal{D}^1_{(n-1)}| + |\mathcal{D}^2_{(n-1)}|)$. This number cannot be more than the maximum number of dichotomies on $S_{(n-1)}$ where no subset of size $(k+1)$ is shattered. Hence

$$|\mathcal{D}^1_{(n-1)}| + |\mathcal{D}^2_{(n-1)}| \leq \mathbb{B}(n-1, k+1)$$

Also $|\mathcal{D}^2_{(n-1)}| \leq \mathbb{B}(n-1, k)$. Otherwise $\mathcal{D}^2_{(n-1)}$ would shatter at least one subset of size k. However since $\mathcal{D}_n$ contains both $\mathcal{D}_{(n-1)} + (\mathbf{x}_0 \to +1)$ and $\mathcal{D}_{(n-1)} + (\mathbf{x}_0 \to -1)$, it would follow that $\mathcal{D}_n$ shatters a subset of size $(k+1)$. This violates our definition of $\mathbb{B}(n, k+1)$. Putting the two together we get the recurrence

$$\mathbb{B}(n, k+1) \leq \mathbb{B}(n-1, k+1) + \mathbb{B}(n-1, k)$$

Proof of the fact that this recurrence implies the upper bound in the theorem is left as Exercise 18.

The second part of the lemma is consequence of the following sequence of inequalities.

$$\begin{aligned}
\sum_{i=0}^{k} \binom{n}{i} &\leq \left(\frac{n}{k}\right)^k \sum_{i=0}^{k} \binom{n}{i} \left(\frac{k}{n}\right)^i \\
&\leq \left(\frac{n}{k}\right)^k \sum_{i=0}^{n} \binom{n}{i} \left(\frac{k}{n}\right)^i \\
&= \left(\frac{n}{k}\right)^k \left(1 + \frac{k}{n}\right)^n \\
&\leq \left(\frac{n}{k}\right)^k e^k = \left(\frac{en}{k}\right)^k
\end{aligned}$$

■

An easy corollary to the above theorem is that for any function class $\mathcal{F}$ with a finite VC-dimension $\mathsf{d}_{\mathcal{F}}$

$$\mathcal{G}_{\mathcal{F}}(n) = O(n^{\mathsf{d}_{\mathcal{F}}}) \tag{2}$$

Exercise 16 *Show by induction that in particular for finite VC-Dimension*

$$\mathcal{G}_{\mathcal{F}}(n) \leq n^{\mathsf{d}_{\mathcal{F}}} + 1$$

The relationship between the Growth function, VC-Dimension and n is illustrated in Figure 4.

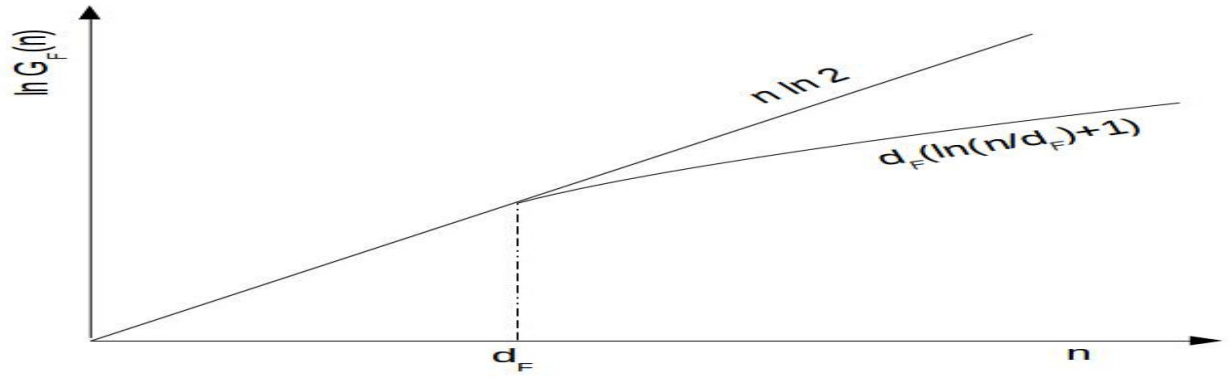

Figure 4: Growth Function, VC-Dimension and Sample Size

Exercise 17 *Show that* $\mathbb{B}(n,1) = 1$ *and* $\mathbb{B}(1,k) = 2$ *for any* $k > 1$.

Exercise 18 *Assuming the recurrence*

$$\mathbb{B}(n,k) \leq \mathbb{B}(n-1,k) + \mathbb{B}(n-1,k-1)$$

show by induction on $(n+k)$ *that*

$$\mathbb{B}(n,k) \leq \sum_{i=0}^{k-1} \binom{n}{i}$$

Exercise 19 *Let* $N_{\mathcal{H}}(\mathcal{D}_n)$ *be the number of dichotomies generated by the hypothesis class* $\mathcal{H}$ *on a training sample* $\mathcal{D}_n$. *Define the* VC-Entropy $H(n)$ *of* $\mathcal{H}$ *as*

$$H(n) = E_{\mathcal{D}_n}\left[\ln N_{\mathcal{H}}(\mathcal{D}_n)\right]$$

and the Annealed VC-Entropy *of* $\mathcal{H}$ *as*

$$H_{ann}(n) = \ln E_{\mathcal{D}_n}\left[N_{\mathcal{H}}(\mathcal{D}_n)\right]$$

Show that

$$H(n) \leq H_{ann}(n) \leq \ln \mathcal{G}_{\mathcal{H}}(n)$$

***Hint**: Use Jensen's Inequality.*

Exercise 20 *Let* $\{\mathcal{H}_1, \ldots, \mathcal{H}_k\}$ *be* k *hypothesis classes each of VC-dimension* $\mathsf{d}_{\mathcal{H}}$ *and let* $\mathcal{F} = \cup_{i=1}^{k} \mathcal{H}_i$. *Show the following:*

1. $\mathsf{d}_{\mathcal{F}} < k.(1 + \mathsf{d}_{\mathcal{H}})$
2. $2^t \geq k.t^{(\mathsf{d}_{\mathcal{H}}+1)} \Rightarrow \mathsf{d}_{\mathcal{F}} \leq t$

3 Generalization Bounds

In this section we present the fundamental theorem of learning — that

Theorem 3 ***VC Theorem****: Given any target binary labelling function function $f : \mathcal{D} \rightarrow \pm 1$ to be learnt on a domain $\mathcal{D}$, any input probability distribution $p_d(\boldsymbol{x})$ on $\mathcal{D}$, any hypothesis class $\mathcal{H}$ with growth function $\mathcal{G}_{\mathcal{H}}$, any $0 < \epsilon < 1$, then*

$$\forall h \in \mathcal{H} : \; Pr\left[|R(h) - R_e(h)| > \epsilon \right] \leq 4.\mathcal{G}_{\mathcal{H}}(2n).e^{-\frac{1}{8}\epsilon^2 n}$$

where the training dataset $\mathcal{D}_n$ is of size n and the probability is over all possible training datasets $\mathcal{D}_n$.

Proof: We will show this by showing that

$$Pr\left[\sup_{h \in \mathcal{H}} |R(h) - R_e(h)| > \epsilon\right] \leq 4.\mathcal{G}_{\mathcal{H}}(2n).e^{-\frac{1}{8}\epsilon^2 n}$$

The idea behind the proof has already been explored a little earlier in our proof of the No Free Lunch Theorem — taking a equal sized 'ghost' dataset and using the ghost dataset $\mathcal{D}'_n$ as a proxy for the larger domain. We assume that $\mathcal{D}_n \cap \mathcal{D}'_n = \phi$. Let us introduce the following notation for the rest of the proof — these are typical events corresponding to any given hypothesis h:

$$\begin{aligned}
\Delta(h) &= (|R(h) - R_e(h)| > \epsilon), \quad \Delta^+ = \left(\sup_{h \in \mathcal{H}} |R(h) - R_e(h)| > \epsilon\right) \\
\Delta'(h) &= (|R(h) - R'_e(h)| \leq \frac{\epsilon}{2} \\
\Delta_e(h) &= \left(|R_e(h) - R'_e(h)| > \frac{\epsilon}{2}\right), \quad \Delta_e^+ = \left(\sup_{h \in \mathcal{H}} |R_e(h) - R'_e(h)| > \frac{\epsilon}{2}\right)
\end{aligned}$$

where $R_e(h)$ and $R'_e(h)$ are the empirical risks suffered by the hypothesis h in the training dataset $\mathcal{D}_n$ and the ghost dataset $\mathcal{D}'_n$ respectively. $R(h)$ is the true risk of h across the domain. The 'plan' behind the proof is as follows:

1. We will bound $\Delta(h)$ in terms of $\Delta_e(h)$. Note that it is easier to analyze $\Delta_e(h)$ since this is purely a combinatorial problem involving two finite datasets — the original dataset $\mathcal{D}_n$ and the ghost dataset $\mathcal{D}'_n$. Specifically we show that:

   $$Pr(\Delta^+) \leq 2Pr(\Delta_e^+) \tag{3}$$

   Note that the probability on the LHS is over all possible choices of $\mathcal{D}_n$ and that on the RHS is over all possible choices of $\mathcal{D}_{2n}$.

2. Imagine $\mathcal{D}_n$ and $\mathcal{D}_n$ have been produced by first sampling an i.i.d dataset $\mathcal{D}_{2n}$ of size $2n$ from $\mathcal{D}$ and selecting n points from $\mathcal{D}_{2n}$ uniformly at random to get $\mathcal{D}_n$. Then $\mathcal{D}'_n$ is taken as $\mathcal{D}_{2n} - \mathcal{D}_n$. We then bound $Pr(\Delta_e^+)$ in terms of the conditional probability $Pr(\Delta_e(h) \mid \mathcal{D}_{2n})$. Specifically we show that

   $$Pr(\Delta_e^+) \leq G_{\mathcal{H}}(2n). \sup_{\mathcal{D}_{2n}} \sup_{h \in \mathcal{H}} Pr(\Delta_e(h) \mid \mathcal{D}_{2n}) \tag{4}$$

3. We then use a variant of Hoeffding's inequality that applies to sampling without replacement to bound this conditional probability. We show that:

$$\sup_{\mathcal{D}_{2n}} \sup_{h \epsilon \mathcal{H}} Pr(\Delta_e(h) \mid \mathcal{D}_{2n}) \leq 2e^{-\frac{1}{8}\epsilon^2 n} \tag{5}$$

Putting Equations 3, 4 and 5 together proves the theorem. We prove Equations 3, 4 and 5 in separate lemmas below. ■

Lemma 2 *Given a hypothesis class $\mathcal{H}$, dataset $\mathcal{D}_n$ of size n and a ghost dataset $\mathcal{D}'_n$*

$$\left(1 - 2e^{-\frac{1}{2}\epsilon^2 n}\right) Pr(\Delta^+) \leq Pr(\Delta_e^+)$$

Proof: Starting from the RHS of the inequality we have

$$\begin{aligned} Pr(\Delta_e^+) &\geq Pr(\Delta_e^+ \wedge \Delta^+) \\ &= Pr(\Delta^+).Pr(\Delta_e^+ \mid \Delta^+) \end{aligned}$$

The second term of the last expression can be estimated as follows: let h^* be a hypothesis that satisfies the event Δ^+ on which the probability is conditioned. That is $\Delta(h^*)$ is true (i.e., $|R(h^*) - R_e(h^*)| > \epsilon$). Also since $\Delta_e(h^*) \Rightarrow \Delta_e^+$ we have that

$$Pr(\Delta_e^+ \mid \Delta^+) \geq Pr\left(\Delta_e(h^*) \mid \Delta^+\right)$$

Therefore

$$Pr(\Delta_e^+) \geq Pr(\Delta^+).Pr\left(\Delta_e(h^*) \mid \Delta^+\right)$$

It is easy to verify that given $\Delta(h^*)$, $\Delta'(h^*) \Rightarrow \Delta_e(h^*)$. Hence

$$Pr\left(\Delta_e(h^*) \mid \Delta^+\right) \geq Pr\left(\Delta'(h^*) \mid \Delta^+\right)$$

Therefore

$$Pr(\Delta_e^+) \geq Pr(\Delta^+).Pr\left(\Delta'(h^*) \mid \Delta^+\right) \tag{6}$$

The probability on the right is over all possible choices of $\mathcal{D}_n$ and potentially h^* may be different for different $\mathcal{D}_n$s. However for any $\mathcal{D}_n$, h^* is independent of $\mathcal{D}'_n$ and hence Heoffding's inequality applies to $Pr\left(\Delta'(h^*) \mid \Delta^+\right)$. Therefore

$$Pr\left(\Delta'(h^*) \mid \Delta^+\right) \geq (1 - 2e^{-\frac{1}{2}\epsilon^2 n})$$

For the bound in the VC-Theorem to be non-trivial we require that

$$e^{-\frac{1}{2}\epsilon^2 n} < e^{-\frac{1}{8}\epsilon^2 n} < \frac{1}{4} \Rightarrow (1 - 2e^{-\frac{1}{2}\epsilon^2 n}) \geq \frac{1}{2}$$

This along with Equation 6 establishes the lemma. ■

Lemma 3

$$Pr(\Delta_e^+) \leq G_{\mathcal{H}}(2n). \sup_{\mathcal{D}_{2n}} \sup_{h \epsilon \mathcal{H}} Pr(\Delta_e(h) \mid \mathcal{D}_{2n})$$

Proof: Let $h_1, \ldots, h_{\mathcal{G}_{\mathcal{H}}(2n)} \in \mathcal{H}$ be hypotheses in $\mathcal{H}$ representing the equivalence classes of hypotheses where each class consists of the all the hypotheses that generate identical labellings of $\mathcal{D}_{2n}$. Note that $\mathcal{G}_{\mathcal{H}}(2n)$ is the growth function of $\mathcal{H}$ representing the largest number of distinct dichotomies that can be generated on $\mathcal{D}_{2n}$ by $\mathcal{H}$. The following series of inequalities prove the lemma.

$$\begin{aligned}
Pr(\Delta_e^+) &= \sum_{\mathcal{D}_{2n}} Pr(\mathcal{D}_{2n}).Pr(\Delta_e^+ \mid \mathcal{D}_{2n}) \\
&\leq \sup_{\mathcal{D}_{2n}} Pr(\Delta_e^+ \mid \mathcal{D}_{2n}) \\
&= \sup_{\mathcal{D}_{2n}} Pr\left(\bigvee_{i=1}^{\mathcal{G}_{\mathcal{H}}(2n)} (\Delta_e(h_i) \mid \mathcal{D}_{2n})\right) \\
&\leq \sup_{\mathcal{D}_{2n}} \sum_{i=1}^{\mathcal{G}_{\mathcal{H}}(2n)} Pr\left(\Delta_e(h_i) \mid \mathcal{D}_{2n}\right) \\
&\leq \mathcal{G}_{\mathcal{H}}(2n). \sup_{\mathcal{D}_{2n}} \sup_{h \in \mathcal{H}} Pr\left(\Delta_e(h) \mid \mathcal{D}_{2n}\right)
\end{aligned}$$

■

Lemma 4

$$\sup_{\mathcal{D}_{2n}} \sup_{h \in \mathcal{H}} Pr(\Delta_e(h) \mid \mathcal{D}_{2n}) \leq 2e^{-\frac{1}{8}\epsilon^2 n}$$

Proof: We take recourse to a variant of Hoeffding's Inequality [**hoeffding**] that applies to sampling without replacement — this violates the i.i.d assumption. Hoeffding's inequality states that for a sequence of random samples $X_1, \ldots, X_n$ chosen without replacement from a distribution with expected value μ

$$Pr\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > t\right) \leq 2e^{-2t^2 n}$$

Given a $\mathcal{D}_{2n}$, we sample n points out of $\mathcal{D}_{2n}$ without replacement to form $\mathcal{D}_n$. Let $X_1, \ldots, X_{2n}$ represent the losses for a given hypothesis h at each of the points in $\mathcal{D}_{2n}$ and without loss of generality let $X_1, \ldots, X_n \in \mathcal{D}_n$. Then $X_{n+1}, \ldots, X_{2n} \in \mathcal{D}'_n$. The following are now easy to check:

$$R_e(h) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad R'_e(h) = \frac{1}{n}\sum_{i=n+1}^{2n} X_i, \quad \mu = \frac{1}{2n}\sum_{i=1}^{2n} X_i = \frac{R_e(h) + R'_e(h)}{2}$$

Using these in the Hoeffding's inequality along with $t = \frac{\epsilon}{4}$ along with the observation that we have not made any assumptions about h or $\mathcal{D}_{2n}$ so far, gives

$$\forall \mathcal{D}_{2n} \subset \mathcal{D}, \ \forall h \in \mathcal{H}, \quad Pr(\Delta_e(h) \mid \mathcal{D}_{2n}) \leq 2e^{-\frac{1}{8}\epsilon^2 n}$$

The claim of the lemma is an immediate consequence of the above. ■

4 Implications of the VC-Theorem

The VC-Theorem has several implications to learning. We will explore some of these in this section. The statement of the theorem can be rewritten in several different ways.

1. For every hypothesis $h \in \mathcal{H}$, given an i.i.d sample of size n with probability at least $(1-\delta)$
$$R(h) \leq R_e(h) + \sqrt{\frac{8}{n} \ln \frac{4.\mathcal{G}_{\mathcal{H}}(2n)}{\delta}}$$

2. The empirical risk of every hypothesis in a hypothesis class for which $\lim_{n\to\infty} \frac{\ln \mathcal{G}_{\mathcal{H}}(n)}{n} = 0$, will converge to its true risk as $n \to \infty$. This also means hypotheses classes with infinite VC-dimension are not PAC learnable. Therefore the ERM principle is consistent if and only if
$$\lim_{n\to\infty} \frac{\ln \mathcal{G}_{\mathcal{H}}(n)}{n} = 0$$

3. Given any $0 < \epsilon, \delta < 1$, for i.i.d samples of size n where
$$n \geq \frac{8}{\epsilon^2} \ln \left(\frac{4.\mathcal{G}_{\mathcal{H}}(2n)}{\delta} \right)$$
the following bound holds:
$$Pr\left(\sup_{h\in\mathcal{H}} |R(h) - R_e(h)| > \epsilon \right) \leq \delta$$

Exercise 21 *Show that the bound in item 3 above is consistent, i.e.,*
$$n_2 \geq n_1 \text{ and } n_1 \geq \frac{8}{\epsilon^2} \ln \left(\frac{4.\mathcal{G}_{\mathcal{H}}(2n_1)}{\delta} \right) \Rightarrow n_2 \geq \frac{8}{\epsilon^2} \ln \left(\frac{4.\mathcal{G}_{\mathcal{H}}(2n_2)}{\delta} \right)$$

The VC-Theorem shows that for any hypothesis class $\mathcal{H}$ with a finite VC-Dimension, every hypothesis in the class will generalize. For the ERM inductive principle it is easy to show that if $h_n^* \in \mathcal{H}$ is the empirical risk minimizing hypothesis and $h^* \in \mathcal{H}$ is the hypothesis that minimizes the risk globally, then
$$\lim_{n\to\infty} Pr\left(|R(h^*) - R_e(h_n^*)| > \epsilon \right) = 0$$
This is from the observation that $R_e(h_n^*) \leq R_e(h^*) \leq R(h^*) \leq R(h_n^*)$.

Recall the definition of the VC-Entropy $H(n)$ of a hypothesis class $\mathcal{H}$ as in Exercise 19. Vapnik and Chervonenkis had actually shown also that the ERM principle is consistent if and only if
$$\lim_{n\to\infty} \frac{H(n)}{n} = 0$$
However this is not a constructive bound since the VC-Entropy depends on the input distribution. The growth function $\mathcal{G}_{\mathcal{H}}(n)$ on the other hand is independent of the input distribution.

Exercise 22 *Show that for the class $\mathcal{H}$ of perceptron (hyperplane) classifiers in d dimensions, with high probability*

$$\sup_{h\in\mathcal{H}} |R(h) - R_e(h)| = O\left(\sqrt{\frac{d}{n}\ln n}\right)$$

Exercise 23 *A universal bound such as the VC-bounds is expected to be very loose. The rate at which learning converges will typically be much better than that guaranteed by the VC-bounds. Carry out the following experiment using your Perceptron code to see the gap between guaranteed bounds and what you observe in practice.*

1. *Fix a value for the dataset size n, the dimensionality of the problem d and an arbitrary distribution $\rho(\mathbf{x})$ over $\mathcal{R}^d$.*
2. *Plot the distribution of ϵ from the theoretical bounds.*
3. *Generate n data points according to $\rho(\mathbf{x})$ and compute a perceptron separator for the same.*
4. *Measure the error (ϵ) committed by the perceptron on a test set generated according to $\rho(\mathbf{x})$.*
5. *Repeat the above two steps and plot the distribution of ϵ for the perceptrons computed.*
6. *Compare this plot of the ϵ-distribution against the theoretical plot.*

Repeat this experiment for multiple values of n and observe the trend.

5 Bias-Variance Tradeoff

Consider a learning algorithm $\mathcal{A}$ trying to learn a target function $f : \mathcal{D} \to \mathcal{R}$ over a domain $\mathcal{D}$ that outputs a hypothesis $h_{\mathcal{D}_n} : \mathcal{D} \to \mathcal{R}$ when presented with a training dataset $\mathcal{D}_n$. The expected risk (across all possible training datasets) incurred by the algorithm, for a square error loss function is

$$\begin{aligned} E_{\mathcal{D}_n}\left[R(h_{\mathcal{D}_n})\right] &= E_{\mathcal{D}_n}\left[E_{\boldsymbol{x}}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\ &= E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}) - f(\boldsymbol{x}))^2\right]\right] \\ &= E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}))^2 + (f(\boldsymbol{x}))^2 - 2f(\boldsymbol{x})h_{\mathcal{D}_n}(\boldsymbol{x})\right]\right] \end{aligned}$$

Let $g(\boldsymbol{x}) : \mathcal{D} \to \mathcal{R}$ be a function defined as

$$g(\boldsymbol{x}) = E_{\mathcal{D}_n}\left[h_{\mathcal{D}_n}(\boldsymbol{x})\right]$$

The value of $g(\boldsymbol{x})$ is simply the expected value of the value output by the hypotheses produced by $\mathcal{A}$, across all possible training datasets. Note that $g(\boldsymbol{x})$ may not necessarily

belong to $\mathcal{H}$. Also observe that $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are independent of $\mathcal{D}_n$. With these we can now extend the above equalities to

$$\begin{aligned} E_{\mathcal{D}_n}\left[R(h_{\mathcal{D}_n})\right] &= E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}))^2\right] + (f(\boldsymbol{x}))^2 - 2f(\boldsymbol{x})g(\boldsymbol{x})\right] \\ &= E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}))^2\right] - (g(\boldsymbol{x}))^2 + (f(\boldsymbol{x}))^2 - 2f(\boldsymbol{x})g(\boldsymbol{x}) + (g(\boldsymbol{x}))^2\right] \\ &= E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right]\right] + E_{\boldsymbol{x}}\left[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right] \end{aligned}$$

The term $E_{\boldsymbol{x}}\left[E_{\mathcal{D}_n}\left[(h_{\mathcal{D}_n}(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right]\right]$ is called the *variance* and the second term $E_{\boldsymbol{x}}\left[(f(\boldsymbol{x}) - g(\boldsymbol{x}))^2\right]$ is called the *bias*. Therefore the earlier expression can now be written as

$$E_{\mathcal{D}_n}\left[R(h_{\mathcal{D}_n})\right] = \textbf{Variance} + \textbf{Bias}$$

This represents an important trade-off in learning. The 'bias' term is the expected error incurred by the algorithm across the domain. The variance' term is a measure of how sensitive the algorithm is to changes in the training dataset — this is the variance of the hypothesis produced by $\mathcal{A}$ around the 'average' hypothesis. For instance in an ERM setting

1. A learning algorithm with a 'complex' hypothesis class will arrive at a hypothesis that whose empirical risk is close to zero. That would mean that the 'average' hypothesis $g(\boldsymbol{x})$ will be close to the target function $f(\boldsymbol{x})$ implying that the bias is low. On the other hand the algorithm becomes very sensitive to the training data - small changes in the training dataset can result in large changes in the output hypothesis. The variance of the hypothesis around its 'average' is large.

2. A learning algorithm with a simple hypothesis class (the class of constant hypotheses for example) will hardly produce anything different even when presented with very different looking training data sets. However the algorithm would have a large bias and would produce hypotheses that deviate significantly from the target function.

We can now reinterpret the learning curve for the hypothesis $h_{\mathcal{D}_n}$ produced by $\mathcal{A}$ in terms of the bias and variance. See Figure 5.

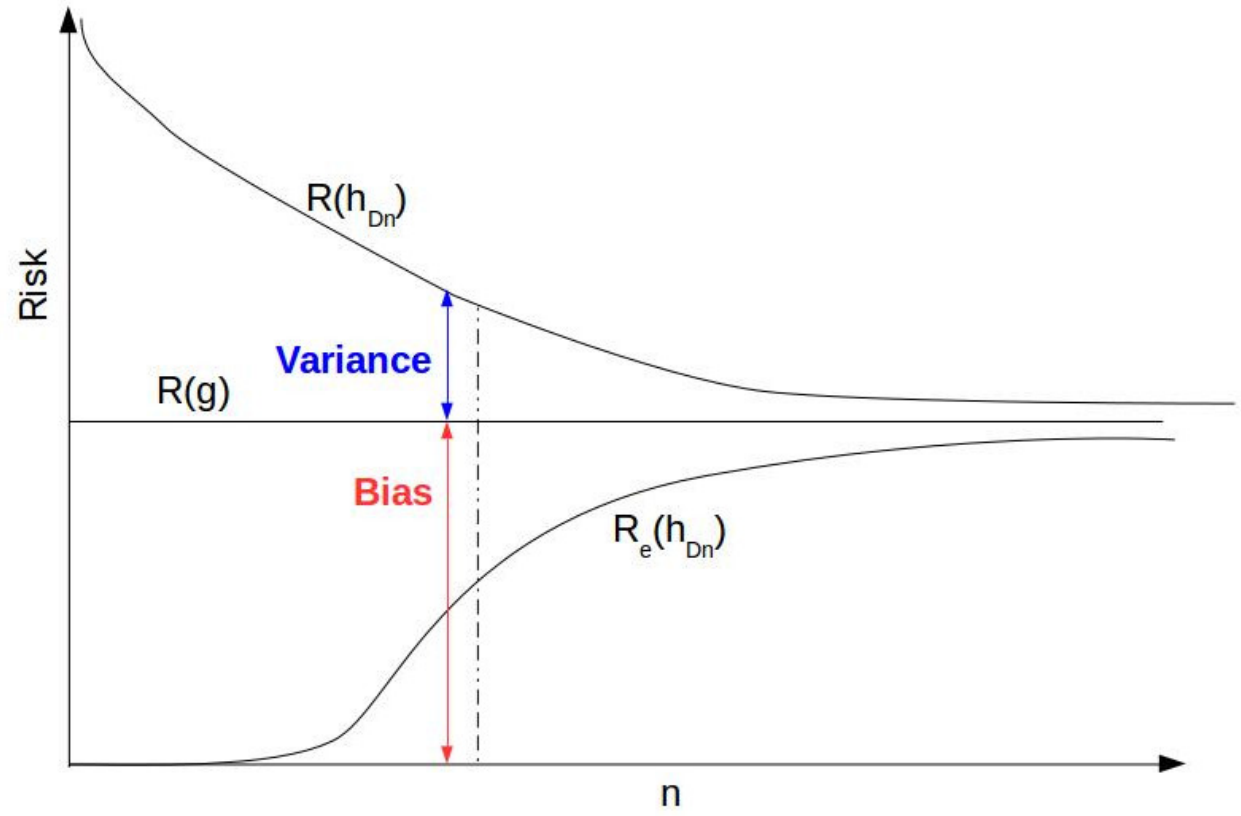

Figure 5: Bias Variance Tradeoff