TTIC 31120: Learning Theory **October 8, 2018**

Lecture 3: VC Theory, No Free Lunch and the Fundamental Theorem of Statistical Lea

Lecturer: Nati Srebro *Scribes: Shuyan Huang*

In this lecture, we will introduce VC dimension as a formal way to characterize the number of "parameters" of a hypothesis class. And instead of cardinality, we will give a convergence bound of ERM based on VC dimension. We will also show some examples of calculating VC dimension and introduce half space representation theorem. Finally, we will present statistical no free lunch theorem and give a proof of a stronger statement.

3.1 Cardinality and Sample Complexity

We have given a upper bound of sample complexity based on cardinality:

$$m_{\mathcal{H}}(\epsilon,\delta) \leq m_{ERM,\mathcal{H}}(\epsilon,\delta) \leq O(\frac{\log|\mathcal{H}| + \log 1/\delta}{\epsilon^2})$$

This shows that all finite hypothesis classes are PAC learnable. However, it remains to be seen whether it is also a necessary condition. If not, there can be infinite classes that are learnable. Also, is the bound on $m_{\mathcal{H}}$ always tight? Are all classes of the same cardinality equally complex?

Example 1. The following two hypothesis classes are of the same cardinality $|\mathcal{H}| = 2^{100}$:

$$\mathcal{X} = \{1,...,100\},\ \mathcal{H} = \{\pm 1\}^{\mathcal{X}}$$
$$\mathcal{X} = \{1,...,2^{100} \approx 10^{30}\},\ \mathcal{H} = \{[[x \leq \theta]] | \theta \in 1...2^{100}\}$$

But apparently the later class is easier to learn since each sample gives more information. So the sample complexity of the later class should be smaller. This implies that there might be a tighter bound on the sample complexity.

3.2 The Growth Function

We first give a bound based on the growth function. The growth function of a hypothesis class is defined as follows:

Definition 1. *(Growth Function) For a given sample set C of length m, the growth function $\Gamma_{\mathcal{H}}(C)$ is the number of different functions from set C to {0,1} that can be obtained by restricting H to C. And for length m, the growth function $\Gamma_{\mathcal{H}}(m)$ is the maximum of the growth functions of all sets of length m. For $C = (x_1, x_2, ..., x_m) \in \mathcal{X}_m$:*

$$\Gamma_{\mathcal{H}}(C) = |\{(h(x_1), h(x_2), ..., h(x_m)) \in \{\pm 1\}^m | h \in \mathcal{H}\}|$$

For m:

$$\Gamma_{\mathcal{H}}(m) = \max_{c \in \mathcal{X}^m} \Gamma_{\mathcal{H}}(C)$$

Again let's see the previous two examples

Example 2. $\mathcal{X} = \{1,...,100\}$, $\mathcal{H} = \{\pm 1\}^{\mathcal{X}}$. When $m \leq 100$, there are at most 2^m different combinations of labels that can be obtained by some $h \in \mathcal{H}$. When $m > 100$, there are repetition in samples, but at most 2^{100} behaviors can be obtained from $\mathcal{H}$. Therefore, $\Gamma_{\mathcal{H}}(m) = min(2^m, 2^{100})$.

Example 3. $\mathcal{X} = \{1,...,2^{100} \approx 10^{30}\}$, $\mathcal{H} = \{[[x \leq \theta]] | \theta \in 1...2^{100}\}$. When $m < 2^{100}$, in the best case, the samples are m different points from $\{1,...,2^{100}\}$, we can obtain $m+1$ different label combinations by putting θ into the $m+1$ intervals between points, before the smallest point and after the largest point. Since there are only 2^{100} different choices of θ, there are at most 2^{100} behaviors that can be obtained from $\mathcal{H}$. Therefore, $\Gamma_{\mathcal{H}}(m) = min(m+1, 2^{100})$.

3.3 Uniform Convergence using the Growth Function

Instead of cardinality, we now give a uniform convergence bound using the growth function:

Theorem 4. *For any hypothesis class $\mathcal{H}$ and any mcD,*

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\forall_{h\in\mathcal{H}}, |L_{\mathcal{D}}(h) - L_S(h)| \leq 4\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log 2/\delta}{m}}\right] \geq 1-\delta$$

From this we can conclude the following bounds:
For any $\mathcal{H}$ and any $\mathcal{D}$, $\forall^{\delta}_{S\sim\mathcal{D}^m}$,

$$L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + 4\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log 2/\delta}{m}}$$

and

$$L_{\mathcal{D}}(\hat{h}) \leq \inf_{h\in\mathcal{H}} L_{\mathcal{D}}(\hat{h}) + 8\sqrt{\frac{\log|\Gamma_{\mathcal{H}}(2m)| + \log 2/\delta}{m}}$$

3.4 Shattering and VC Dimension

Definition 2. *(Shattering) $C = \{x_1, ..., x_m\}$ is shattered by $\mathcal{H}$ if $\Gamma_{\mathcal{H}}(C) = 2^m$, i.e. we can get all 2^m behaviors:*

$$\forall_{y_1,...,y_m\in\pm 1},\ \exists_{h\in\mathcal{H}}\ s.t.\ \forall_i h(x_i) = y_i$$

Definition 3. *(VC dimension) The VC dimension of $\mathcal{H}$ is the largest number of points that can be shattered by $\mathcal{H}$:*

$$VCdim(\mathcal{H}) = \max m\ s.t.\ \Gamma_{\mathcal{H}}(C) = 2^m$$

Example 5. As a special case, when $\mathcal{H}$ is infinite and $\forall_m \Gamma_{\mathcal{H}}(C) = 2^m$, then $VCdim(\mathcal{H}) = \inf$.

Example 6.

$$\mathcal{X} = \{1, ..., 100\},\ \mathcal{H} = \{\pm 1\}^{\mathcal{X}}$$

We have shown that when the sample size is larger than 100, $\mathcal{H}$ can't shatter all the samples. $VCdim = 100$.

Example 7. (Discrete Threshold)

$$\mathcal{X} = \{1, ..., 2^{100} \approx 10^{30}\},\ \mathcal{H} = \{[[x \leq \theta]] | \theta \in 1...2^{100}\}$$

If a sample has 2 points $x_1 < x_2$, we can never get the behavior $x_1 = -1, x_2 = 1$. So we can not shatter more than 1 point, $VCdim = 1$

Example 8. (Continuous Threshold)

$$\mathcal{X} = \mathbb{R},\ \mathcal{H} = \{[[x \leq \theta]] | \theta \in \mathbb{R}\}$$

Still, we can not get the behavior in the previous example, $VCdim = 1$

Example 9. (Intervals)

$$\mathcal{X} = \mathbb{R},\ \mathcal{H} = \{h_{a,b} = [[a \leq x \leq b]] | a, b \in \mathbb{R}\}$$

We can shatter any two points, but with three points $x_1 < x_2 < x_3$, we can't realize $x_1 = 1, x_2 = -1, x_3 = 1$. So $VCdim = 2$

Example 10. (Axis Aligned Rectangles)

$$\mathcal{X} = \mathbb{R}^2, \mathcal{H} = \{h_{a_1,a_2,b_1,b_2} = [[a_1 \leq x_1 \leq a_2 \text{ and } b_1 \leq x_2 \leq b_2]] | a_1, a_2, b_1, b_2 \in \mathbb{R}\}$$

We can shatter 1,2,3 points. Some sets of 4 points can't be shattered:

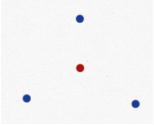

Figure 3.1: 4 points that can't be shattered by axis aligned rectangles

When the blue points are +1 and the red point is -1, this can't be realized by any axis aligned rectangle. But according to the definition, as long as we find 4 points that can be shattered by $\mathcal{H}$, $VCdim(\mathcal{H}) \geq 4$. We find:

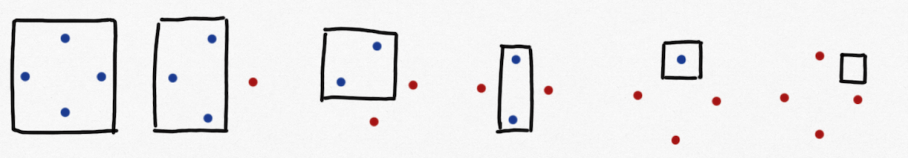

Figure 3.2: 4 points that can be shattered by axis aligned rectangles

In Figure 3.2, blue points are given label +1 and red points are given label -1. Figure 3.2 shows how different combinations of labels can be realized by axis aligned rectangles. This shows that $VCdim \geq 4$. To show any 5 points can't be shattered, let's consider any 5 points $S = p_1, p_2, p_3, p_4, p_5$. From S, take a leftmost point, a rightmost point, a lowest point, and a highest point. Without loss of generality, let p_5 be the point that was not selected(there might be several points that are not selected but we can just choose one). Now, define the labeling (1,1,1,1,0). It is impossible to obtain this labeling by an axis aligned rectangle. Indeed, such a rectangle must contain p_1, p_2, p_3, p_4, but in this case the rectangle must contain p_5 as well, because its coordinates are within the intervals defined by the selected points. So, S is not shattered by $\mathcal{H}$, and therefore $VCdim(\mathcal{H}) = 4$.

To give a uniform convergence bound by VC dimension, we first build the relationship between VC dimension and growth function, then use Theorem 4.

Lemma 11. *(Sauer-Shelah-VC Lemma) If $VCdim(\mathcal{H}) = D$, then:*

$$\Gamma_{\mathcal{H}}(m) \leq \sum_{i=0}^{D} \binom{m}{i} \leq (\frac{em}{D})^D$$

Figure 3.3 plots the growth function and 2^m. When the sample size m is smaller than VC dimension, the growth function grows exponentially with m. After the sample size m reaches VC dimension, the growth function grows much more slowly(polynomially) with m.

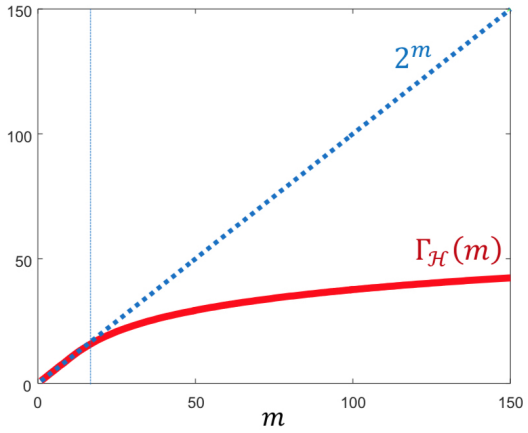

Figure 3.3: Plotting growth function and 2^m

From Sauer Lemma, we get:

$$\log|\Gamma_{\mathcal{H}}(2m)| \leq \log\left(\frac{em}{VCdim}\right)^{VCdim} \leq O(VCdim \cdot \log m)$$

Plugging this into Theorem 4, we obtain:

Theorem 12.

$$\forall^{\delta}_{S\sim\mathcal{D}^m},\ L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{VCdim(\mathcal{H})\log m + \log 1/\delta}{m}}\right)$$

With a very complex proof, this can be improved to:

$$\forall^{\delta}_{S\sim\mathcal{D}^m},\ L_{\mathcal{D}}(\hat{h}) \leq L_S(\hat{h}) + O\left(\sqrt{\frac{VCdim(\mathcal{H}) + \log 1/\delta}{m}}\right)$$

We now show more examples of VC dimension:

Example 13. (Circles in $\mathbb{R}^2$)

$$\mathcal{H} = \{h_{c,r} = [[\| x - c \| \leq r]] | c \in \mathbb{R}^2, r \in \mathbb{R}\}$$

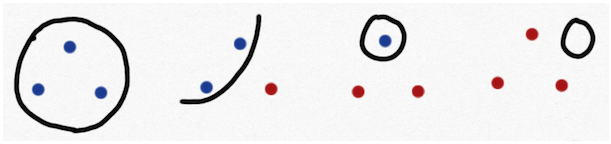

Figure 3.4: 3 points that can be shattered by circles

As shown in Figure 3.4, where blue points are given +1 label and red points are given -1 label, circles can shatter 3 points. So $VCdim(\mathcal{H}) \geq 3$.

Example 14. (Circles and their complement)

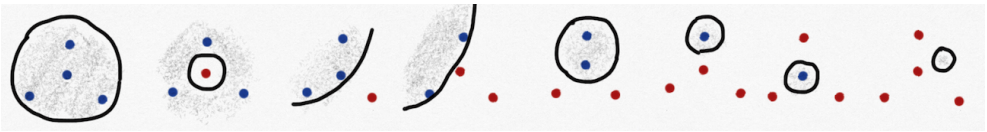

Figure 3.5: 3 points that can be shattered by circles and their complement

As shown in Figure 3.5, where blue points are given +1 label, red points are given -1 label, and the regions of +1 label are in shadow, circles and their complement can shatter 4 points. So $VCdim(\mathcal{H}) \geq 4$.

Example 15. (Circles around origin)

$$\mathcal{H} = \{h_r = [[\| x \| \leq r]] | r \in \mathbb{R}\}$$

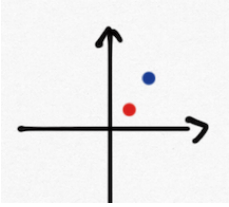

Figure 3.6: 2 points that can't be shattered by circles around origin

Apparently $\mathcal{H}$ can shatter 1 point. As shown in Figure 3.5, any 2 points can't be shattered by circles around origin, because $\mathcal{H}$ can't produce the labeling where the point closer to the origin is -1 while the other is +1. So $VCdim(\mathcal{H}) = 1$.

Example 16. (Axis aligned ellipses)

$$\mathcal{H} = \{h_{c,r} = [[\frac{(x[1]-c[1])^2}{r[1]^2} + \frac{(x[2]-c[2])^2}{r[2]^2} \leq 1]] | c, r \in \mathbb{R}^2\}$$

Can shatter 4 points, so $VCdim(\mathcal{H}) \geq 4$.

Example 17. (General ellipses) Can shatter 5 points, so $VCdim(\mathcal{H}) \geq 5$.

Example 18. (Homogeneous half spaces)

$$\mathcal{H}_\phi = \{[[\langle w, \phi(x) \geq 0]] | w \in \mathbb{R}^d\}, \qquad \phi : \mathcal{X} \to \mathbb{R}^d$$

$\mathcal{H}_\phi$ can shatter the d standard basis $e_1, ..., e_d$ by choosing $w = (y_1, y_2, ..., y_d)$.
Now we prove the claim that $\mathcal{H}_\phi$ can't shatter any set of $d+1$ points.

Proof. For any $d+1$ points $x_1, ..., x_{d+1}$, there must be some linear dependency:$\sum_i a_i x_i = 0$. Consider the labeling $y_i = sign(a_i)$(arbitrary for $a_i = 0$). If we can shatter, there's a predictor w s.t.

$$sign(\langle w, x_i \rangle) = y_i = sign(a_i)$$

This implies that when $a_i \neq 0$,

$$a_i \langle w, x_i \rangle > 0$$

We then have:

$$\langle w, \sum_i a_i x_i \rangle = \sum_i a_i \langle w, x_i \rangle > 0$$

Which contradicts with the linear dependency:

$$\langle w, \sum_i a_i x_i \rangle = 0$$

□

Since $\mathcal{H}_\phi$ can shatter d points but not $d+1$ points, $VCdim(\mathcal{H}_\phi) = d$

Based on the above result, we have the following half space representation theorem:

Theorem 19. *(half space representation) For a hypothesis class $\mathcal{H}$, if there exists $\phi : \mathcal{X} \to \mathbb{R}^D$ s.t.*

$$\mathcal{H} \subseteq \mathcal{H}_\phi,$$

i.e. every hypothesis $h \in \mathcal{H}$ can be written as $h(x) = sign(\langle w, \phi(x)\rangle)$ for some $w_h \in \mathbb{R}^D$, then $VCdim(\mathcal{H}) \leq D$.

Example 20. Non-homogeneous half-spaces over $\mathbb{R}^d$ can be represented by $(d+1)$-dimension homogeneous half-spaces with $\tilde{\phi}(x) = [\phi(x), 1]$

Example 21. A circle can be represented by a 4-dimension half-space:

$$\begin{aligned} h_{a,r}(x) &= [[(x[1]-a[1])^2 + (x[2]-a[2])^2 \leq r^2]] \\ &= [[x[1]^2 - 2a[1]x[1] + a[1]^2 + x[2]^2 - 2a[2]x[2] + a[2]^2 \leq r^2]] \\ &= sign((-x[1]^2 - x[2]^2) + 2a[1]x[1] + 2a[2]x[2] + (r^2 - a[1]^2 - a[2]^2)1) \\ &= sign(\langle w_{a,r}, \phi(x)\rangle) \end{aligned}$$

where $\phi(x) = (-x[1]^2 - x[2]^2, x[1], x[2], 1)$ and $w_{a,r} = (1, 2a[1], 2a[2], r^2 - a[1]^2 - a[2]^2)$. According to Theorem 19, $VCdim \leq 4$. We showed in Example 13 that the lower bound of VC dimension is 3, so why isn't the upper bound tight? Notice that $w[1]$ is restricted to 1. In fact, if we allow $w[1] < 0$, we get circles and their complement(including half-spaces, which can be thought of as infinite radius circles), and half space representation gives a tight bound on VC dimension($VCdim = 4$).

Example 22. (Axis-aligned ellipses and their complement)

$$\phi(x) = (x[1]^2, x[2]^2, x[1], x[2], 1)$$

$VCdim \leq 5$

Example 23. (Conic cuts (including all ellipses))

$$\phi(x) = (x[1]^2, x[2]^2, x[1]x[2], x[1], x[2], 1)$$

$VCdim \leq 6$

Example 24. (Degree-k polynomials over $\mathbb{R}^2$)

$$\begin{aligned} \phi(x) = (&x[1]^k, x[1]^{k-1}x[2]^1, ..., x[1]^1x[2]^{k-1}, x[2]^k, \\ &x[1]^{k-1}, x[1]^{k-2}x[2]^1, ..., x[1]^1x[2]^{k-2}, x[2]^{k-1}, \\ &x[1]^{k-2}, ... \\ &x[1]^2, x[1]x[2], x[2]^2, \\ &x[1], x[2], 1) \in \mathbb{R}^{(k+1)(k+2)/2} \end{aligned}$$

$VCdim \leq (k+1)(k+2)/2$

Example 25. (Degree-k polynomials over $\mathbb{R}^d$)

$$\phi(x) \in \mathbb{R}^{\binom{d+k-1}{k}}$$

$VCdim \leq O(d^k)$

In the previous examples, the VC dimension happened to equal the number of parameters defining the hypothesis class, which makes us to think about this question: is VC dimension always equal to the number of parameters? The answer is yes when we are dealing with most "normal" hypothesis classes. But it isn't always the case. Look at the following counter example:

Example 26.

$$\mathcal{X} = \mathbb{R} \qquad \mathcal{H} = \{h_{\theta,v}(x) = sign(sin(vx + \theta)) | v, \theta \in \mathbb{R}\}$$

We now prove that $VCdim(\mathcal{H}) = \infty$:

Proof. Consider the infinite set of points $\{x_i = 10^{-i}\}_{i=1,2,...}$. Any labeling $y_1, y_2, ...$ can be obtained by $\theta = 0$ and:

$$v = \pi \left(1 - \sum_{i=1}^{\infty} \frac{y_i}{2x_i}\right)$$

□

Back to learning, we have shown that the sample complexity of ERM can be given by cardinality and VC dimension:

$$m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{\log|\mathcal{H}| + \log 1/\delta}{\epsilon^2}\right)$$

$$m_{ERM,\mathcal{H}}(\epsilon, \delta) = O\left(\frac{VCdim(\mathcal{H}) + \log 1/\delta}{\epsilon^2}\right)$$

We can still ask these questions: Can a class with infinite VC-dimension be learnable? Can the sample complexity be lower than the VC dimension? We have shown what is learnable, but we haven't shown what is not learnable. Are there any hypothesis classes that are not learnable? Can we learn the class $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ of all possible predictors?

3.5 Statistical No Free Lunch

In this section will prove that there's no learning rule that can succeed on all source distributions.

Theorem 27. *(Statistical No Free Lunch) For any domain $\mathcal{X}$ of size $|\mathcal{X}|$ and any learning rule A, there exists a source distribution $\mathcal{D}$ with $\mathbb{P}_{x,y\sim\mathcal{D}}[f(x) = y] = 1$ for some $f : \mathcal{X} \to \{\pm 1\}$, such that for $m < \frac{|X|}{2}$,*

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{4}$$

and so w.p. $\geq 1/7$, $L_{\mathcal{D}}(A(S)) \geq 1/8$.
For an infinite domain $\mathcal{X}$, for any learning rule A and any sample size m, there exists a source distribution and f as above such that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[L_{\mathcal{D}}(A(S))] \geq \frac{1}{4}$$

Instead of proving Statistical No Free Lunch Theorem, we prove a stronger statement:

Theorem 28. *For a finite domain $\mathcal{X}$, $\mathcal{Y} = \{\pm 1\}$, any learning rule A and any sample size m,*

$$\frac{1}{2} - \frac{m}{2|\mathcal{X}|} \leq \mathbb{E}_f \mathbb{E}_{S\sim\mathcal{U}_f^m}[L_{\mathcal{U}_f}(A(S))] \leq \frac{1}{2} + \frac{m}{2|\mathcal{X}|}$$

The first expectation is taken over a uniform distribution of $f : \mathcal{X} \to \mathcal{Y}$ where for each x set, w.p. $1/2$, $f(x) = \pm 1$, independent of all other values. The second expectation is taken over samples of size m from $\mathcal{U}_f$, which is the source distribution s.t. x is uniform over $\mathcal{X}$ and $y = f(x)$ with probability one.

Before proving this theorem, we define:

Definition 4. *S is consistent with f if* $\forall_{(x_i,y_i)\in S} f(x_i) = y_i$. *S is self-consistent if it is consistent with some f (i.e. if* $x_i = x_j$ *then* $y_i = y_j$*).*

Proof. For any learning rule A, and any self-consistent sample S:

$$\begin{aligned}
&\mathbb{E}_f[L_{\mathcal{U}_f}(A(S))|S \text{ cons with } f] \\
&= \mathbb{E}_f[\frac{1}{|\mathcal{X}|}\sum_x [[A(S)(x) \neq f(x)]]|S \text{ cons with } f] \\
&= \frac{1}{|\mathcal{X}|}\left(\sum_{x\in S}\mathbb{P}_f[A(S)(x)\neq f(x)] + \sum_{x\notin S}\mathbb{P}_f[A(S)(x)\neq f(x)]\right)
\end{aligned}$$

For x that appears in sample S, the best a learning rule can do is to get all of these x right, and the worst case is when a learning rule gets all of these x wrong. For x that doesn't appear in sample S, learning rules have no information of it, so by taking expectation over all consistent f, all learning rule can get half of these x right (since f divides the labels into ± 1 with equal probability). So $\frac{1}{2} - \frac{m'}{2|\mathcal{X}|} \leq \mathbb{E}_f[L_{\mathcal{U}_f}(A(S))|S \text{ cons with } f] \leq \frac{1}{2} + \frac{m'}{2|\mathcal{X}|}$, where $m' = |\{x|x \in S\}|$. And so:

$$\begin{aligned}
\mathbb{E}_f\mathbb{E}_{S\sim\mathcal{U}_f^m}[L_{\mathcal{U}_f}(A(S))] &= \mathbb{E}_f\mathbb{E}_{self-cons\ S}[L_{\mathcal{U}_f}(A(S))|S \text{ cons with } f] \\
&= \mathbb{E}_{self-cons\ S}\mathbb{E}_f[L_{\mathcal{U}_f}(A(S))|S \text{ cons with } f]
\end{aligned}$$

$$\frac{1}{2} - \frac{m}{2|\mathcal{X}|} \leq \mathbb{E}_f\mathbb{E}_{S\sim\mathcal{U}_f^m}[L_{\mathcal{U}_f}(A(S))] \leq \frac{1}{2} + \frac{m}{2|\mathcal{X}|}$$

□

Finally, we can use Statistical No Free Lunch Theorem to show that the class $\mathcal{H} = \mathcal{Y}^{\mathcal{X}}$ is not learnable for infinite $\mathcal{X}$; For finite $\mathcal{X}$, sample complexity is $m_{\mathcal{Y}^{\mathcal{X}}}(1/8, 6/7) \geq \Omega(|\mathcal{X}|) = \Omega(VCdim(\mathcal{H})) = \Omega(\log|\mathcal{H}|)$