CPSC 532D — 6. GROWTH FUNCTIONS AND VC DIMENSION

Danica J. Sutherland

University of British Columbia, Vancouver

Fall 2023

So far, we've mainly talked about logistic regression. We proved some bounds that we obtain nearly the optimal value of the logistic loss, but we haven't actually said anything yet about 0-1 loss (i.e. accuracy). How can we handle binary classifiers, in terms of accuracy?

First, recall a couple of key things from our last set of notes:

$$\ell \circ \mathcal{H} = \{z \mapsto \ell(h, z) : h \in \mathcal{H}\} \qquad \mathcal{F}|_S = \left\{\big(f(z_1), \dots, f(z_m)\big) : f \in \mathcal{F}\right\} \subseteq \mathbb{R}^m$$

$$\operatorname{Rad}(V) = \frac{1}{m} \mathop{\mathbb{E}}_{\sigma \sim \operatorname{Unif}(\pm 1)^m} \sup_{v \in V} \sigma \cdot v$$

$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le 2 \mathop{\mathbb{E}}_{S \sim \mathcal{D}^m} \operatorname{Rad}((\ell \circ \mathcal{H})|_S)$$

$$\text{Talagrand: if } \ell(h, (x, y)) = l_y(h(x)) \text{ for } \rho\text{-Lipschitz } l_y,\ \operatorname{Rad}((\ell \circ \mathcal{H})|_S) \le \rho \operatorname{Rad}(\mathcal{H}|_{S_x})$$

$$\text{If } \ell \in [a, b],\ \Pr_{S \sim \mathcal{D}^m}\left(\sup_{h \in \mathcal{H}} L_{\mathcal{D}}(h) - L_S(h) \le \mathbb{E} \sup_{h \in \mathcal{H}} [L_{\mathcal{D}}(h) - L_S(h)] + (b - a)\sqrt{\frac{\log \frac{1}{\delta}}{2m}} \right) \ge 1 - \delta.$$

We're going to focus for now on binary classifiers, i.e. h that output a binary label, not necessarily a continuous one like in logistic regression. (Maybe "inside" h there's a continuous value that then gets thresholded, but we won't be modeling that directly.)

1 ZERO-ONE LOSS

If $h(x) \in \{-1, 1\}$ and $y \in \{-1, 1\}$, then the 0-1 loss is

$$l_y(\hat{y}) = \begin{cases} 0 & \hat{y} = y \\ 1 & \hat{y} \ne y. \end{cases}$$

This isn't a function on $\mathbb{R}$, so applying Talagrand's lemma is a little weird. The trick is, though: for computing the loss, we can just extend the function l_y to $\mathbb{R}$ in any way at all, and the loss will be exactly the same – it just doesn't care what l_y does for *other* values of $\hat{y}$.

So, let's just pick a Lipschitz function on $\mathbb{R}$ that agrees at the points we need, by linear interpolation:

$$l_y(\hat{y}) = \begin{cases} 0 & y\hat{y} \ge 1 \\ \frac{1}{2} - \frac{1}{2} y\hat{y} & 0 \le y\hat{y} \le 1 \\ 1 & y\hat{y} \le 1. \end{cases}$$

For more, visit https://cs.ubc.ca/~dsuth/532D/23w1/.

This has $\|l_y\|_{\mathrm{Lip}} = \frac{1}{2}|y| = \frac{1}{2}$. So, for any $\mathcal{H}_{-1,1}$ of binary classifiers mapping to $\{-1, 1\}$,

$$\mathrm{Rad}((\ell_{0-1} \circ \mathcal{H}_{-1,1})|_{S}) \le \frac{1}{2}\mathrm{Rad}(\mathcal{H}_{-1,1}|_{S_x}). \tag{1}$$

Note that, since Rad is "scale-sensitive", this very much depended on the choice to do $\{-1, 1\}$ classifiers. If we have $\{0, 1\}$ classifiers, we can either choose a slightly different function (which would be 1-Lipschitz), or note that we can convert from a $\{0, 1\}$ classifier to a $\{-1, 1\}$ classifier by taking $2h - 1$ and use basic properties of Rademacher complexity to see

$$\mathrm{Rad}(\mathcal{H}_{-1,1}|_{S_x}) = \mathrm{Rad}((2\mathcal{H}_{0,1} - 1)|_{S_x}) = 2\,\mathrm{Rad}(\mathcal{H}_{0,1}|_{S_x}),$$

so that

$$\mathrm{Rad}((\ell_{0-1} \circ \mathcal{H}_{0,1})|_{S}) \le \mathrm{Rad}(\mathcal{H}_{0,1}|_{S_x}).$$

2 FINITE SETS

How do we bound $\mathrm{Rad}(\mathcal{H}|_{S_x})$ for binary classifiers?

One major way is to note that, for binary classifiers,

$$\mathcal{H}|_{S_x} = \{(h(x_1), \dots, h(x_n)) : h \in \mathcal{H}\} \subseteq \{0, 1\}^m$$

– and so it can't be too big. There are only 2^m possible bitvectors of behaviour on the particular set S_x, *even if $\mathcal{H}$ is infinite*. In fact, there may be many fewer possible things that $\mathcal{H}$ is able to do on this particular S_x.

So, let's first try bounding the Rademacher complexity of an arbitrary finite set based on its size.

LEMMA 1. *If* V *is finite and* $\|v\| \le B$ *for all* $v \in V$, *then*

$$\mathrm{Rad}(V) \le \frac{B}{m}\sqrt{2\log|V|}.$$

Proof. We have

$$\mathrm{Rad}(V) = \mathop{\mathbb{E}}_{\sigma} \max_{v \in V} \sum_{i=1}^{m} \frac{\sigma_i v_i}{m}.$$

Considering any one v for now, $\sum_{i=1}^{m} \sigma_i v_i$ is a random variable (depending on σ). It has mean zero, and since σ_i is $\mathcal{SG}(\frac{1-(-1)}{2}) = \mathcal{SG}(1)$ by Hoeffding's lemma, $v_i\sigma_i/m$ is $\mathcal{SG}(|v_i|/m)$. The $v_i\sigma_i/m$ for each i are independent of one another, so this means

$$\sum_i \frac{v_i\sigma_i}{m} \in \mathcal{SG}\left(\sqrt{\sum_i (|v_i|/m)^2}\right) = \mathcal{SG}(\|v\|/m) \subseteq \mathcal{SG}(B/m).$$

We now want to find the expected max of these $|V|$ random variables. Each is mean zero and $\mathcal{SG}(B/m)$; they're dependent, since they all use the same σ, but that's okay. Lemma 2 handles exactly this situation. □

LEMMA 2. *Let* $T_1, \dots, T_n$ *be zero-mean random variables that are each* $\mathcal{SG}(\sigma)$, *which are*

not *necessarily independent. Then*

$$\mathbb{E}\left[\max_{i\in[n]} T_i\right] \le \sigma\sqrt{2\log(n)}.$$

Proof. This is Assignment 2, Question 2.4. □

For binary classifiers mapping to ±1, $|h(x)| = 1$ so $\left\|h|_{S_x}\right\| = \sqrt{m}$. Thus

COROLLARY 3. *For binary classifiers mapping to* $\{-1, 1\}$, $\mathrm{Rad}(\mathcal{H}|_{S_x}) \le \sqrt{\frac{2}{m}\log\left|\mathcal{H}|_{S_x}\right|}$.

For binary classifiers mapping to $\{0, 1\}$, it's half of that, by our usual scaling-and-translating conversion.

Thus (1) and Theorem 8 from last time give that for binary classifiers and zero-one loss,

You might want to check for yourself that this same equation holds whether $\mathcal{H}$ maps to $\{0, 1\}$ or $\{-1, 1\}$, or indeed any other two-element set.

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}}[L_\mathcal{D}(\mathcal{H}) - L_S(\mathcal{H})] \le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sqrt{\frac{2}{m}\log\left|\mathcal{H}|_{S_x}\right|} \tag{2}$$

$$\Pr_{S\sim\mathcal{D}^m}\left(L_\mathcal{D}(\mathrm{ERM}_\mathcal{H}(S)) - \inf_{h\in\mathcal{H}} L_\mathcal{D}(h) \le \mathop{\mathbb{E}}_{S\sim\mathcal{D}^m}\sqrt{\frac{2}{m}\log\left|\mathcal{H}|_{S_x}\right|} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}}\right) \ge 1-\delta. \tag{3}$$

Using just that $\left|\mathcal{H}|_{S_x}\right| \le |\mathcal{H}|$, this becomes that with probability at least $1-\delta$

$$L_\mathcal{D}(\mathrm{ERM}_\mathcal{H}(S)) - \min_{h\in\mathcal{H}} L_\mathcal{D}(h) \le \sqrt{\frac{2}{m}\log|\mathcal{H}|} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}},$$

which is very similar to the much more direct bound from lecture 2 of

$$L_\mathcal{D}(\mathrm{ERM}_\mathcal{H}(S)) - \min_{h\in\mathcal{H}} L_\mathcal{D}(h) \le \sqrt{\frac{2}{m}\log\frac{|\mathcal{H}|+1}{\delta}} \le \sqrt{\frac{2}{m}\left[\log|\mathcal{H}| + \log\frac{2}{\delta}\right]}.$$

3 GROWTH FUNCTION

The bound $\left|\mathcal{H}|_{S_x}\right| \le |\mathcal{H}|$ is potentially very, very loose, though. For instance, we know the left-hand side can't be more than 2^m, even if the right-hand side is infinite.

Plugging in that 2^m bound would only get us that $\mathrm{Rad}(\mathcal{H}|_{S_x}) \le \sqrt{2\log 2} \approx 1.18$, which is not very interesting since the generalization gap is trivially at most 1! But, when $\left|\mathcal{H}|_{S_x}\right| = o(2^m)$, this is far more interesting.

DEFINITION 4. The *growth function* $\Gamma_\mathcal{H}(m)$ of a hypothesis class $\mathcal{H}$ is given by

$$\Gamma_\mathcal{H}(m) = \sup_{x_1,\dots,x_m\in\mathcal{X}} \left|\mathcal{H}|_{(x_1,\dots,x_m)}\right|.$$

By definition, $\left|\mathcal{H}_{S_x}\right| \le \Gamma_\mathcal{H}(m)$ for any S_x of size m; thus for binary classifiers with zero-one loss, we immediately know that the expected worst-case generalization gap is at most $\sqrt{\frac{2}{m}\log\Gamma_\mathcal{H}(m)}$.

Note that we've dropped all dependence on the particular distribution $\mathcal{D}$; this is now a purely combinatorial notion.

It's sometimes possible to compute growth functions directly – you'll do this for a simple case on assignment 3 – but it's usually going to be much easier to bound it with the VC dimension.

4 VC DIMENSION

DEFINITION 5. A hypothesis class $\mathcal{H}$ is said to *shatter* a set $S_x \subseteq \mathcal{X}$ if it can achieves all possible labellings of S_x, i.e. $\left|\mathcal{H}|_{S_x}\right| = 2^m$.

The letters VC are after Vladimir Vapnik and Alexey Chervonenkis, who developed this theory starting in the 60s in the Soviet Union (well before the definition of PAC learning); the English translation of the first key paper is [VC71].

DEFINITION 6. The *VC dimension* of $\mathcal{H}$ is the size of the largest set $\mathcal{H}$ can shatter:

$$\mathrm{VCdim}(\mathcal{H}) = \max\left(\{m \geq 0 : \Gamma_{\mathcal{H}}(m) = 2^m\}\right).$$

If $\mathcal{H}$ can shatter unboundedly large sets, we say its VC dimension is infinite.

It turns out that we can bound the growth function in terms of the VC dimension: $\Gamma_{\mathcal{H}}(m) = \mathcal{O}(m^{\mathrm{VCdim}(\mathcal{H})})$, which then gives us that the expected worst-case generalization gap is $\tilde{\mathcal{O}}(\sqrt{2\,\mathrm{VCdim}(\mathcal{H})/m})$. We'll see this later; let's first explore how to compute the VC dimension for some different $\mathcal{H}$.

4.1 *Examples of computing VC dimension*

It will be useful for all of our examples below to note that if you can't shatter any set of size m, you also can't shatter any set of size $m' > m$: if you could, then by definition you could shatter any size-m subset of the larger set.

4.1.1 Threshold functions

Let $h_a : \mathbb{R} \to \{0, 1\}$ be a threshold function $h_a(x) = \mathbb{1}(x \geq a)$, and let $\mathcal{H} = \{h_a : a \in \mathbb{R}\}$.

We can shatter any set of size 1, but for VC dimension we only have to show that we can shatter one *particular set of that size.*

To start: we can shatter, say, $S_x = \{0\}$, because $h_{-1}(0) = 1$ and $h_1(0) = 0$. Thus $\mathrm{VCdim}(\mathcal{H}) \geq |S_x| = 1$.

But we can't shatter any set S_x of size $|S_x| \geq 2$. Let $a, b \in S_x$ with $a < b$. We can't get $h(a) = 1$ and $h(b) = 0$ with the same $h \in \mathcal{H}$, since all $h \in \mathcal{H}$ are nondecreasing. Thus no S_x of size 2 be shattered, and so $\mathrm{VCdim}(\mathcal{H}) < 2$.

Thus $\mathrm{VCdim}(\mathcal{H}) = 1$.

4.1.2 Circles

This is like the problem from A2 Q1, but not necessarily centred at the origin. Don't use this technique on A2. :)

For $\mathcal{X} = \mathbb{R}^2$, consider $\mathcal{H} = \{h_{r,c} : r > 0, c \in \mathbb{R}^2\}$ with $h_{r,c}(x) = \mathbb{1}(\|x - c\| \leq r\}$, the set of indicator functions of circles.

We can shatter any set of size two, since we can draw a circle that includes both points, one that includes either point, or one that includes neither point.

A bunch of these examples are easier to see if you draw them out! But I'm not taking the time to draw the diagrams with TikZ this time – sorry. Try drawing them yourself, or check the board pictures.

We can also shatter *some* sets of size three, since if we put them in an equilateral triangle we can pick out none, or any one, two, or all three points. (If we put the three points in a line, we can't pick out the two edges but not the middle – but that's okay, VC dimension is about *the largest* set you can shatter.)

Claim: we cannot shatter any set of size four, and so $\mathrm{VCdim}(\mathcal{H}) = 3$. If we think of the points as lying roughly in a rectangle, then we can't pick out opposite corners without including at least one of the other points. (Ideally you'd formalize this argument, but let's not do that now.)

4.1.3 Homogeneous linear threshold functions in $\mathbb{R}^2$

Let $\mathcal{X} = \mathbb{R}^2$ and consider $\mathcal{H} = \{x \mapsto \operatorname{sgn}(w \cdot x) : w \in \mathbb{R}^2\}$: hyperplanes passing through the origin. We're now using $\mathcal{Y} = \{-1, 1\}$, and we're going to define a function sgn which is like the sign except that $\operatorname{sgn}(0) = 1$ – yeah, that sucks but so do all the other options. If you want to stick to $\mathcal{Y} = \{0, 1\}$, then instead use $\mathbb{1}(w \cdot x \geq 0)$; that's much nicer to write down, but more annoying to work with.

We can shatter at least some sets of size 2: e.g. $\{(-1, 1), (1, 1)\}$, we can put the hyperplane along the x-axis to get both the same sign, or put it in along the y-axis to get them with opposite signs.

We can't shatter any sets of size 3. If the convex hull of the points contains the origin, then we can't get them all with the same sign; if the hull doesn't contain the origin, then we can't label them like $(1, 0, 1)$.

A convex hull *of a set is the smallest convex set containing the original set:* $\operatorname{conv}(V) = \{\alpha v + (1-\alpha)v' : v, v' \in V, \alpha \in [0,1]\}$. *If you have some points in* $\mathbb{R}^2$, *you draw straight lines connecting the "outside" points to include all the points.*

So homogenous 2-d linear threshold functions have VC dimension 2.

4.1.4 Homogeneous linear threshold functions in $\mathbb{R}^d$

PROPOSITION 7. *Let* $\mathcal{H} = \{x \mapsto \operatorname{sgn}(w \cdot x) : w \in \mathbb{R}^d\}$. *Then* $\operatorname{VCdim}(\mathcal{H}) = d$.

Proof. We can shatter a set of size d: take the set $\{e_1, \dots, e_d\}$ for e_i the ith standard basis vector, i.e. the one-hot vector with a 1 in the ith position and 0 everywhere else. Then we can achieve an arbitrary labeling $(y_1, \dots, y_d) \in \{0,1\}^d$ by setting $w_i = y_i$: we get $w \cdot e_i = y_i$.

Now, let $x_1, \dots, x_{d+1}$ be a set of $d+1$ points in $\mathbb{R}^d$. Then they can't be linearly independent: there must be some $\alpha_1, \dots, \alpha_{d+1}$ such that $\sum_{i=1}^{d+1} \alpha_i x_i = 0$, with not all the α_i zero. Let $\mathcal{I}_+ = \{i \in [d+1] : \alpha_i > 0\}$, $\mathcal{I}_0 = \{i \in [d+1] : \alpha_i = 0\}$, and $\mathcal{I}_- = \{i \in [d+1] : \alpha_j < 0\}$.

Now, if $\mathcal{H}$ can shatter $\{x_1, \dots, x_{d+1}\}$, we can ask it to assign 1 to the x_i with $i \in \mathcal{I}_+ \cup \mathcal{I}_0$, and -1 to the x_i with $i \in \mathcal{I}_-$. Then we'd have

$$0 = w \cdot 0 = w \cdot \sum_{i=1}^{d+1} (\alpha_i x_i) = \sum_{i \in \mathcal{I}_+} \underbrace{\alpha_i}_{>0} \underbrace{w \cdot x_i}_{\geq 0} + \sum_{i \in \mathcal{I}_0} \underbrace{\alpha_i}_{0} \, w \cdot x_i + \sum_{i \in \mathcal{I}_-} \underbrace{\alpha_i}_{<0} \underbrace{w \cdot x_i}_{<0} .$$

I claim that the sum on the right-hand side is strictly positive, meaning we've shown $0 < 0$, a contradiction; thus $\mathcal{H}$ cannot shatter $\{x_1, \dots, x_{d+1}\}$. This is easiest to see if $\mathcal{I}_-$ is nonempty: those terms will all be strictly positive. It will also be positive if there are any points in $\mathcal{I}_+$ with $w \cdot x_i > 0$. Otherwise, the only case left is if $w \cdot x_i = 0$ for all $i \in \mathcal{I}_+$ and $\mathcal{I}_- = \{\}$, meaning that this w labels all of the data points as positive. Since $\mathcal{I}_- = \{\}$, we must then have $\sum_{i \in \mathcal{I}_+} \alpha_i x_i = 0$. Now, find some $\tilde{w}$ that labels all these points as negative, $\tilde{w} \cdot x_i < 0$ for all $i \in \mathcal{I}_+$; this must be possible if the set is shattered. Then we'd have

[SSBD] misses analyzing this case :(, since they just pretend $w \cdot x = 0$ *is impossible.*

$$0 = \tilde{w} \cdot 0 = \tilde{w} \cdot \left(\sum_{i \in \mathcal{I}_+} \alpha_i x_i \right) = \sum_{i \in \mathcal{I}_+} \underbrace{\alpha_i}_{>0} \underbrace{\tilde{w} \cdot x_i}_{<0} < 0,$$

a contradiction. Thus $\mathcal{H}$ cannot shatter $\{x_1, \dots, x_{d+1}\}$. □

4.1.5 Inhomogeneous linear threshold functions in $\mathbb{R}^d$

What about if we don't enforce that the hyperplane passes through the origin, $\mathcal{H} = \{x \mapsto \operatorname{sgn}(w \cdot x + b) : w \in \mathbb{R}^d, b \in \mathbb{R}\}$?

We could analyze this directly, similarly to what we did above; it's Example 3.12 of [MRT] if you want to see it.

But we can also reduce to the set of homogeneous linear classifiers: if we have d-dimensional data, we can model that as homogeneous linear classifiers on $(d+1)$-dimensional data with an extra "dummy feature" that's always 1. The weight w_0 corresponding to that feature will just be the offset b.

Using this reduction, we can see:

PROPOSITION 8. *For $x \in \mathbb{R}^d$, $\operatorname{VCdim}\left(\left\{x \mapsto w \cdot x + b : w \in \mathbb{R}^d, b \in \mathbb{R}\right\}\right) = d + 1$.*

Proof. First, we can shatter the set $\{0, e_1, \dots, e_d\}$, which has size $d+1$, like before. We set $w_0 = y_0/2$ and $w_i = y_i$; the $y_0/2$ only affects the sign if all the other weights are "off", i.e. only on the 0 vector.

Also, we can't shatter any set of size $d+2$. If we could, then there would be $d+2$ vectors in $\mathbb{R}^{d+1}$ shattered by the class of homogeneous thresholds; but that class has VC dimension $d+1$ by Proposition 7, so that's not possible. □

4.2 *Growth function bounds in terms of VC: Sauer-Shelah*

As mentioned before, we're going to show that $\Gamma_{\mathcal{H}}(m)$ is $\mathcal{O}(m^{\operatorname{VCdim}(\mathcal{H})})$. Remember that for $m \le \operatorname{VCdim}(\mathcal{H})$, we know that $\Gamma_{\mathcal{H}}(m) = 2^m$; this means that $\Gamma_{\mathcal{H}}$ always grows exponentially up to some point, then drops off to just polynomial growth.

This e is $\exp(1) \approx 2.7$.

COROLLARY 9. *If $m \ge d = \operatorname{VCdim}(\mathcal{H})$, then $\Gamma_{\mathcal{H}}(m) \le \left(\frac{em}{d}\right)^d$.*

Plugging into (2) and (3) gives

THEOREM 10. *Let $\mathcal{H}$ be a class of binary classifiers with $\operatorname{VCdim}(\mathcal{H}) = d$, and use the zero-one loss. For any $m \ge d$, we have that*

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^m} \sup_{h\in\mathcal{H}}[L_{\mathcal{D}}(\mathcal{H}) - L_S(\mathcal{H})] \le \sqrt{\frac{2d}{m}[\log m + 1 - \log d]}$$

$$\Pr_{S\sim\mathcal{D}^m}\left(L_{\mathcal{D}}(\mathrm{ERM}_{\mathcal{H}}(S)) - \min_{h\in\mathcal{H}} L_{\mathcal{D}}(h) \le \sqrt{\frac{2d}{m}[\log m + 1 - \log d]} + \sqrt{\frac{2}{m}\log\frac{2}{\delta}}\right) \ge 1 - \delta.$$

When $d \ge 3$, we can replace $\log m + 1 - \log d$ with simply $\log m$ above.

We'll prove Corollary 9 as a corollary to the following:

LEMMA 11 (Sauer-Shelah). *Let $\operatorname{VCdim}(\mathcal{H}) \le d < \infty$. Then $\Gamma_{\mathcal{H}}(m) \le \sum_{i=0}^{d} \binom{m}{i}$.*

Proof of Corollary 9 given Lemma 11. We need to show that $\sum_{i=0}^{d}\binom{m}{i} \le \left(\frac{em}{d}\right)^d$ for $m \ge d$.

We can do this by

$$\begin{aligned}
\sum_{i=0}^{d}\binom{m}{i} &\leq \sum_{i=0}^{d}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} && \text{multiply each term by} \geq 1 \\
&\leq \sum_{i=0}^{m}\binom{m}{i}\left(\frac{m}{d}\right)^{d-i} && \text{add nonnegative terms} \\
&= \left(\frac{m}{d}\right)^{d}\sum_{i=0}^{m}\binom{m}{i}\left(\frac{d}{m}\right)^{i} \\
&= \left(\frac{m}{d}\right)^{d}\left(1+\frac{d}{m}\right)^{m} && \text{binomial theorem} \\
&\leq \left(\frac{m}{d}\right)^{d} e^{d} && 1+x \leq \exp(x). \qquad \square
\end{aligned}$$

Now, we'll actually prove Lemma 11 itself as a corollary to the following result:

LEMMA 12 (Pajor). *For all finite* $\mathrm{S} \subseteq \mathcal{X}$, $\left|\mathcal{H}|_{\mathrm{S}}\right| \leq \left|\{\mathrm{T} \subseteq \mathrm{S} : \mathrm{T} \textit{ is shattered by } \mathcal{H}\}\right|$.

If S_x is shattered, both sides of the inequality are $2^{|\mathrm{S}_x|}$; otherwise, it's not obvious that these things should be related.

Proof of Lemma 11 given Lemma 12. To bound the number of shattered subsets of S in Lemma 12, recall there can't possibly be any with size larger than $d = \mathrm{VCdim}(\mathcal{H})$; the number of sets it can shatter is thus upper-bounded by the number of subsets of S of size at most d, which is just $\sum_{i=0}^{d}\binom{n}{i}$ for $n = |\mathrm{S}|$. □

Proof of Lemma 12. We'll proceed by (strong) induction on $\left|\mathcal{H}|_{\mathrm{S}}\right|$, writing just S instead of S_x for brevity.

Base case: $\left|\mathcal{H}|_{\mathrm{S}}\right| = 1$. For the right-hand side, the empty set is trivially shattered by any $\mathcal{H}$, so the RHS is always at least 1 as well, and the inequality holds.

Inductive case: $\left|\mathcal{H}|_{\mathrm{S}}\right| \geq 2$ and the inequality holds for any T with $\left|\mathcal{H}|_{\mathrm{T}}\right| < \left|\mathcal{H}|_{\mathrm{S}}\right|$. Then, since there are two distinct labelings, there must be at least one point $x \in \mathrm{S}$ that achieves both $h(x) = 1$ and $h'(x) = 0$ for some $h, h' \in \mathcal{H}$. Partition $\mathcal{H}$ into $\mathcal{H}_+ = \{h \in \mathcal{H} : h(x) = 1\}$ and $\mathcal{H}_- = \{h \in \mathcal{H} : h(x) = 0\}$. Now,

$$\left|\mathcal{H}|_{\mathrm{S}}\right| = \left|\mathcal{H}_+|_{\mathrm{S}}\right| + \left|\mathcal{H}_-|_{\mathrm{S}}\right|,$$

since the two produce disjoint labelings on S (they always disagree on x). They also produce fewer labelings than $\mathcal{H}|_{\mathrm{S}}$ itself (there's at least one labeling in each), so we can apply the inductive hypothesis to each.

Defining $\mathrm{Shat}_{\mathcal{H}}(\mathrm{S}) = \{\mathrm{T} \subseteq \mathrm{S} : \mathrm{T} \text{ is shattered by } \mathcal{H}\}$, we've shown that

$$\left|\mathcal{H}|_{\mathrm{S}}\right| \leq \left|\mathrm{Shat}_{\mathcal{H}_+}(\mathrm{S})\right| + \left|\mathrm{Shat}_{\mathcal{H}_-}(\mathrm{S})\right|.$$

Note the right-hand side is exactly, keeping track of the "double-counted" sets,

$$\left|\mathrm{Shat}_{\mathcal{H}_+}(\mathrm{S}) \cup \mathrm{Shat}_{\mathcal{H}_-}(\mathrm{S})\right| + \left|\mathrm{Shat}_{\mathcal{H}_+}(\mathrm{S}) \cap \mathrm{Shat}_{\mathcal{H}_-}(\mathrm{S})\right|;$$

it remains to argue that this is at most $|\mathrm{Shat}_{\mathcal{H}}(S)|$. To see this, first note that $\mathrm{Shat}_{\mathcal{H}_+}(S) \cup \mathrm{Shat}_{\mathcal{H}_-}(S) \subseteq \mathrm{Shat}_{\mathcal{H}}(S)$.

Now, consider a set $T \in \mathrm{Shat}_{\mathcal{H}_+}(S) \cap \mathrm{Shat}_{\mathcal{H}_-}(S)$, i.e. one that's been double-counted. Then note that $T' = T \cup \{x\}$ is not in either $\mathrm{Shat}_{\mathcal{H}_+}(S)$ or $\mathrm{Shat}_{\mathcal{H}_-}(S)$, since these classes cannot shatter $\{x\}$ and so can't shatter a superset of $\{x\}$ either. But $\mathcal{H}$ can shatter T': there's a hypothesis in $\mathcal{H}_-$ to achieve any desired labeling with $h(x) = 0$ (since $T \in \mathrm{Shat}_{\mathcal{H}_-}(S)$), and likewise there's a hypothesis in $\mathcal{H}_+$ for any labeling with $h(x) = 1$. So $T' \in \mathrm{Shat}_{\mathcal{H}}(S)$. Also, each such double-counted T corresponds to a different T', since we're adding the same x to each. Thus

$$\left|\mathrm{Shat}_{\mathcal{H}_+}(S) \cap \mathrm{Shat}_{\mathcal{H}_-}(S)\right| \le \left|\mathrm{Shat}_{\mathcal{H}}(S) \setminus \left(\mathrm{Shat}_{\mathcal{H}_+}(S) \cup \mathrm{Shat}_{\mathcal{H}_-}(S)\right)\right|,$$

and so $\left|\mathcal{H}|_S\right| \le |\mathrm{Shat}_{\mathcal{H}}(S)|$ as desired. □

References

[MRT] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talkwalkar. *Foundations of Machine Learning*. 2nd edition. MIT Press, 2018.

[SSBD] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[VC71] Vladimir N. Vapnik and Alexey Ya. Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities." *Theory of Probability & Its Applications* 16.2 (1971), pages 264–280.