More on VC dimensions

CPSC 532S: Modern Statistical Learning Theory

24 January 2022

cs.ubc.ca/~dsuth/532S/22/

Admin

- A1 solutions are posted (just publicly on the course site)
  - Also see the post-mortem poll on Piazza
  - Grading: probably next week sometime (ICML deadline…)

- A2 will come probably next week + allow (encourage) groups

- No office hours this week (ICML…)
  - Piazza / schedule a meeting if needed

Admin

- A1 solutions are posted (just publicly on the course site)
  - Also see the post-mortem poll on Piazza
  - Grading: probably next week sometime (ICML deadline…)

- A2 will come probably next week + allow (encourage) groups

- No office hours this week (ICML…)
  - Piazza / schedule a meeting if needed

- FYI, now presenting from a different app
  - Going to try out some live scribbles for part of this lecture
  - Will save stuff I write on slides and post after class
  - (Let me know if you hate it)

Last time: shattering / VC dimension

- **Restriction of $\mathscr{H}$ to C** is $\mathscr{H}_C = \left\{ \left(h(c_1), \ldots, h(c_{|C|}) \right) : h \in \mathscr{H} \right\}$
- $\mathscr{H}$ **shatters** $C \subseteq \mathscr{X}$ if $\mathscr{H}_C$ contains all functions from C to $\{0,1\}$: $|\mathscr{H}_C| = 2^{|C|}$
- $\mathrm{VCdim}(\mathscr{H})$ is size of the largest set that $\mathscr{H}$ can shatter, or ∞

- Doesn't need that *all* sets of size VCdim can be shattered – it's **worst-case**
  - There is a C with $|C| = \mathrm{VCdim}$ that can be shattered
  - There is **no** C with $|C| = \mathrm{VCdim} + 1$ that can be shattered

Last time: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$

Last time: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$

- If $\mathrm{VCdim}(\mathscr{H}) = d$:
  - $\mathscr{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}}^{UC} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Last time: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$ saw today
3. $\mathscr{H}$ is agnostic PAC learnable immediate
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$ no free lunch

- If $\mathrm{VCdim}(\mathscr{H}) = d$:
  - $\mathscr{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}}^{UC} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}} \leq \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \leq n_{\mathscr{H}} \leq \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Last time: “Fundamental Theorem of Learning”

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$

saw today

immediate

no free lunch

to show!

- If $\mathrm{VCdim}(\mathscr{H}) = d$:
  - $\mathscr{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le \boxed{n_{\mathscr{H}}^{UC} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]}$
  - $\mathscr{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Last time: "Fundamental Theorem of Learning"

For binary classification with 0-1 loss:

These are all equivalent:

1. $\mathscr{H}$ has the uniform convergence property
2. Any ERM rule agnostically PAC learns $\mathscr{H}$
3. $\mathscr{H}$ is agnostic PAC learnable
4. Any ERM rule PAC learns $\mathscr{H}$
5. $\mathscr{H}$ is PAC learnable
6. $\mathrm{VCdim}(\mathscr{H}) < \infty$

saw today

immediate

no free lunch

to show!

(actually will show something worse today)

- If $\mathrm{VCdim}(\mathscr{H}) = d$:
  - $\mathscr{H}$ has uniform convergence property, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le \boxed{n_{\mathscr{H}}^{UC} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]}$
  - $\mathscr{H}$ is agnostic PAC learnable, $\frac{C_1}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon^2}\left[d + \log\frac{1}{\delta}\right]$
  - $\mathscr{H}$ is PAC learnable, $\frac{C_1}{\varepsilon}\left[d + \log\frac{1}{\delta}\right] \le n_{\mathscr{H}} \le \frac{C_2}{\varepsilon}\left[d\log\frac{1}{\varepsilon} + \log\frac{1}{\delta}\right]$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$

- Will prove in terms of the **growth function**: $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C| = n} |\mathcal{H}_C|$
  - How many *actually different* functions from $\mathcal{H}$ are there on sets of size n?

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \le \varepsilon$ with probability at least $1 - \delta$
- Will prove in terms of the **growth function**: $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C|=n} |\mathcal{H}_C|$
  - How many *actually different* functions from $\mathcal{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n$ for $n \le d$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$

- Will prove in terms of the **growth function**: $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C| = n} |\mathcal{H}_C|$
  - How many *actually different* functions from $\mathcal{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n$ for $n \leq d$

(0,0)
(0,1)
(1,0)
(1,1)

- **Theorem** (SSBD 6.11): $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$

- Will prove in terms of the **growth function**: $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C|=n} |\mathcal{H}_C|$
  - How many *actually different* functions from $\mathcal{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathcal{H}) = d$, then $\tau_{\mathcal{H}}(n) = 2^n$ for $n \leq d$

$\sqrt{\log(2^{2n})} \to \sqrt{2n \log 2}$

- **Theorem** (SSBD 6.11): $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

$\sqrt{\log \mathcal{O}(n^d)} / \sqrt{n} = \dfrac{\sqrt{d\, \mathcal{O}(\log n)}}{\sqrt{n}}$

- **Sauer-Shelah lemma**: when $n \geq d$, $\quad \tau_{\mathcal{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$

- Will prove in terms of the **growth function**: $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C|$
  - How many *actually different* functions from $\mathscr{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathscr{H}) = d$, then $\tau_{\mathscr{H}}(n) = 2^n$ for $n \leq d$

- **Theorem** (SSBD 6.11): $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathscr{H}}(2n))}}{\delta\sqrt{2n}}$

- **Sauer-Shelah lemma**: when $n \geq d$, $\quad \tau_{\mathscr{H}}(n) \leq (en/d)^d = \mathscr{O}\left(n^d\right)$
- Plugging together: uniform convergence when $n \geq 4\dfrac{2d}{(\delta\varepsilon)^2}\log\left(\dfrac{2d}{(\delta\varepsilon)^2}\right) + \dfrac{4d\log(2e/d)}{(\delta\varepsilon)^2}$

Finite VCdim implies uniform convergence

- **Uniform convergence**: $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \varepsilon$ with probability at least $1 - \delta$
- Will prove in terms of the **growth function**: $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C|$
  - How many *actually different* functions from $\mathscr{H}$ are there on sets of size n?
  - If $\mathrm{VCdim}(\mathscr{H}) = d$, then $\tau_{\mathscr{H}}(n) = 2^n$ for $n \leq d$
- **Theorem** (SSBD 6.11): $\sup_{h \in \mathscr{H}} |L_{\mathscr{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathscr{H}}(2n))}}{\delta\sqrt{2n}}$

  We'll come back for a *much* better rate in δ – $\sqrt{\log \frac{1}{\delta}}$ instead of $\frac{1}{\delta}$ – pretty soon
- **Sauer-Shelah lemma**: when $n \geq d$, $\quad \tau_{\mathscr{H}}(n) \leq (en/d)^d = \mathscr{O}\left(n^d\right)$
- Plugging together: uniform convergence when $n \geq 4\dfrac{2d}{(\delta\varepsilon)^2} \log\left(\dfrac{2d}{(\delta\varepsilon)^2}\right) + \dfrac{4d\log(2e/d)}{(\delta\varepsilon)^2}$

Sauer-Shelah lemma

- Independently proved by, at least:
  - Sauer 1972 (to solve a combinatorical problem posed by Erdős)
  - Shelah 1972 (with Perles) as a lemma about “stable models”
  - Perles, later on, in ergodic theory
  - Vapnik+Chervonenkis, also in the 70s, to make VC theory work

- **Sauer-Shelah lemma**: Let $\mathrm{VCdim}(\mathscr{H}) \leq d < \infty$. Then $\tau_{\mathscr{H}}(n) \leq \sum_{i=0}^{d} \binom{n}{i}$.
  - Corollary: for $n \geq d$, $\tau_{\mathscr{H}}(n) \leq \left(\frac{1}{d}en\right)^{d}$

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\, |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathscr{X}$,

  for every $\mathscr{H}$, $\quad |\mathscr{H}_C| \leq \left| \{B \subseteq C : \mathscr{H} \text{ shatters } B\} \right|$

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\, |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathscr{X}$,

  for every $\mathscr{H}$, $\quad |\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$

  - $\mathscr{H}$ doesn't shatter anything bigger than d: RHS $\leq$ (total # of subsets of size $\leq d$)

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathscr{X}$,

  for every $\mathscr{H}$, $\quad |\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$

  - $\mathscr{H}$ doesn't shatter anything bigger than d: RHS $\leq$ (total # of subsets of size $\leq d$)
  - Proof by induction

Proof of Sauer's lemma

- Want to prove $\tau_{\mathcal{H}}(n) = \max_{C \subseteq \mathcal{X}:\ |C|=n} |\mathcal{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathcal{X}$,

  for every $\mathcal{H}$, $\quad |\mathcal{H}_C| \leq \left| \left\{ B \subseteq C : \mathcal{H} \text{ shatters } B \right\} \right|$

  - $\mathcal{H}$ doesn't shatter anything bigger than d: RHS $\leq$ (total # of subsets of size $\leq d$)
  - Proof by induction
  - Base case: $n = 1$, so $C = \{c_1\}$

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathscr{X}$,

  for every $\mathscr{H}$, $\quad |\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$

  - $\mathscr{H}$ doesn't shatter anything bigger than d: RHS $\leq$ (total # of subsets of size $\leq d$)
  - Proof by induction
  - Base case: $n = 1$, so $C = \{c_1\}$
    - If $d = 0$: LHS is 1 ($\mathscr{H}_C$ always $\{0\}$ or $\{1\}$), RHS is just the empty set: 1

Proof of Sauer's lemma

- Want to prove $\tau_{\mathscr{H}}(n) = \max_{C \subseteq \mathscr{X}:\ |C|=n} |\mathscr{H}_C| \leq \sum_{i=0}^{d} \binom{n}{i}$
- Will prove something stronger: for any $C \subseteq \mathscr{X}$,

  for every $\mathscr{H}$, $\quad |\mathscr{H}_C| \leq \left| \{ B \subseteq C : \mathscr{H} \text{ shatters } B \} \right|$

  - $\mathscr{H}$ doesn't shatter anything bigger than d: RHS $\leq$ (total # of subsets of size $\leq d$)
  - Proof by induction
  - Base case: $n = 1$, so $C = \{c_1\}$
    - If $d = 0$: LHS is 1 ($\mathscr{H}_C$ always $\{0\}$ or $\{1\}$), RHS is just the empty set: 1
    - If $d \geq 1$: LHS is 2 ($\mathscr{H}_C = \{0,1\}$), RHS has empty set and C: 2

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$

- Supposing for all C with $|C| < n$, for all $\mathcal{H}$, $|\mathcal{H}_C| \leq \left| \left\{ B \subseteq C : \mathcal{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_n) \in \mathcal{H}_C \right\} = \mathcal{H}_{C'}$

- Supposing for all C with $|C| < n$, for all $\mathcal{H}$, $|\mathcal{H}_C| \leq \left| \left\{ B \subseteq C : \mathcal{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_n) \in \mathcal{H}_C \right\} = \mathcal{H}_{C'}$
    - $|\mathcal{H}_{C'}| \leq \left| \{ B \subseteq C' : \mathcal{H} \text{ shatters } B \} \right| = \left| \{ B \subseteq C : c_1 \notin B, \mathcal{H} \text{ shatters } B \} \right|$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \vee (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}_{C'}$
    - $|\mathscr{H}_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H} \text{ shatters } B \} \right| = \left| \{ B \subseteq C : c_1 \notin B, \mathscr{H} \text{ shatters } B \} \right|$
  - Let $\mathscr{H}' = \left\{ h \in \mathscr{H} : \exists h' \in \mathscr{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{disagree on } c_1 \right\}$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \vee (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}_{C'}$
    - $|\mathscr{H}_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H} \text{ shatters } B \} \right| = \left| \{ B \subseteq C : c_1 \notin B, \mathscr{H} \text{ shatters } B \} \right|$
  - Let $\mathscr{H}' = \left\{ h \in \mathscr{H} : \exists h' \in \mathscr{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{ disagree on } c_1 \right\}$
    - Let $Y_2 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \wedge (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}'_{C'}$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left|\left\{B \subseteq C : \mathscr{H} \text{ shatters } B\right\}\right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$

  - Let $Y_1 = \left\{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \vee (1, y_2, \ldots, y_n) \in \mathscr{H}_C\right\} = \mathscr{H}_{C'}$
    - $|\mathscr{H}_{C'}| \leq \left|\{B \subseteq C' : \mathscr{H} \text{ shatters } B\}\right| = \left|\{B \subseteq C : c_1 \notin B, \mathscr{H} \text{ shatters } B\}\right|$

  - Let $\mathscr{H}' = \left\{h \in \mathscr{H} : \exists h' \in \mathscr{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{disagree on } c_1\right\}$
    - Let $Y_2 = \left\{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \wedge (1, y_2, \ldots, y_n) \in \mathscr{H}_C\right\} = \mathscr{H}'_{C'}$
    - $\mathscr{H}'$ shatters $B \subseteq C'$ iff it shatters $B \cup \{c_1\}$ – happens if $\mathscr{H}$ shatters $B \cup \{c_1\}$

- Supposing for all C with $|C| < n$, for all $\mathcal{H}$, $|\mathcal{H}_C| \leq \left|\left\{B \subseteq C : \mathcal{H} \text{ shatters } B\right\}\right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$

  - Let $Y_1 = \left\{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \vee (1, y_2, \ldots, y_n) \in \mathcal{H}_C\right\} = \mathcal{H}_{C'}$
    - $|\mathcal{H}_{C'}| \leq \left|\{B \subseteq C' : \mathcal{H} \text{ shatters } B\}\right| = \left|\{B \subseteq C : c_1 \notin B, \mathcal{H} \text{ shatters } B\}\right|$

  - Let $\mathcal{H}' = \left\{h \in \mathcal{H} : \exists h' \in \mathcal{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{ disagree on } c_1\right\}$
    - Let $Y_2 = \left\{(y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathcal{H}_C \wedge (1, y_2, \ldots, y_n) \in \mathcal{H}_C\right\} = \mathcal{H}'_{C'}$
    - $\mathcal{H}'$ shatters $B \subseteq C'$ iff it shatters $B \cup \{c_1\}$ – happens if $\mathcal{H}$ shatters $B \cup \{c_1\}$
    - $|\mathcal{H}'_{C'}| \leq \left|\{B \subseteq C' : \mathcal{H}' \text{ shatters } B\}\right| \leq \left|\{B \subseteq C : c_1 \in B, \mathcal{H} \text{ shatters } B\}\right|$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \vee (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}_{C'}$
    - $|\mathscr{H}_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H} \text{ shatters } B \} \right| = \left| \{ B \subseteq C : c_1 \notin B, \mathscr{H} \text{ shatters } B \} \right|$
  - Let $\mathscr{H}' = \left\{ h \in \mathscr{H} : \exists h' \in \mathscr{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{ disagree on } c_1 \right\}$
    - Let $Y_2 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \wedge (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}'_{C'}$
    - $\mathscr{H}'$ shatters $B \subseteq C'$ iff it shatters $B \cup \{c_1\}$ – happens if $\mathscr{H}$ shatters $B \cup \{c_1\}$
    - $|\mathscr{H}'_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H}' \text{ shatters } B \} \right| \leq \left| \{ B \subseteq C : c_1 \in B, \mathscr{H} \text{ shatters } B \} \right|$
  - Have $|\mathscr{H}_C| = |Y_1| + |Y_2|$ since things in Y_2 "show up twice" in $\mathscr{H}_C$

- Supposing for all C with $|C| < n$, for all $\mathscr{H}$, $|\mathscr{H}_C| \leq \left| \left\{ B \subseteq C : \mathscr{H} \text{ shatters } B \right\} \right|$
  - Take $C = (x_1, \ldots, x_n)$, and let $C' = (x_2, \ldots, x_n)$
  - Let $Y_1 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \vee (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}_{C'}$
    - $|\mathscr{H}_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H} \text{ shatters } B \} \right| = \left| \{ B \subseteq C : c_1 \notin B, \mathscr{H} \text{ shatters } B \} \right|$
  - Let $\mathscr{H}' = \left\{ h \in \mathscr{H} : \exists h' \in \mathscr{H} \,.\, h \text{ and } h' \text{ agree on } C', \text{ disagree on } c_1 \right\}$
    - Let $Y_2 = \left\{ (y_2, \ldots, y_n) : (0, y_2, \ldots, y_n) \in \mathscr{H}_C \wedge (1, y_2, \ldots, y_n) \in \mathscr{H}_C \right\} = \mathscr{H}'_{C'}$
    - $\mathscr{H}'$ shatters $B \subseteq C'$ iff it shatters $B \cup \{c_1\}$ – happens if $\mathscr{H}$ shatters $B \cup \{c_1\}$
    - $|\mathscr{H}'_{C'}| \leq \left| \{ B \subseteq C' : \mathscr{H}' \text{ shatters } B \} \right| \leq \left| \{ B \subseteq C : c_1 \in B, \mathscr{H} \text{ shatters } B \} \right|$
  - Have $|\mathscr{H}_C| = |Y_1| + |Y_2|$ since things in Y_2 "show up twice" in $\mathscr{H}_C$
  - So $|\mathscr{H}_C| \leq \left| B \subseteq C : \mathscr{H} \text{ shatters } B \right|$

Proof: corollary to Sauer’s lemma

If $n \geq d$, $\quad \sum_{i=0}^{d} \binom{n}{i}$

Proof: corollary to Sauer's lemma

If $n \geq d$, $\quad \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$ multiply each term by sth ≥ 1

Proof: corollary to Sauer's lemma

If $n \geq d$, $$\sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ multiply each term by sth ≥ 1

$$\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ add nonnegative terms to the sum

Proof: corollary to Sauer's lemma

If $n \geq d$, $$\sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ multiply each term by sth ≥ 1

$$\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ add nonnegative terms to the sum

$$= \left(\frac{n}{d}\right)^{d} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i}$$ rearrange

Proof: corollary to Sauer's lemma

If $n \geq d$,

$$\sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ multiply each term by sth ≥ 1

$$\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$$ add nonnegative terms to the sum

$$= \left(\frac{n}{d}\right)^{d} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i}$$ rearrange

$$= \left(\frac{n}{d}\right)^{d} \left(1 + \frac{d}{n}\right)^{n}$$ binomial theorem

Proof: corollary to Sauer's lemma

If $n \geq d$, $\quad \sum_{i=0}^{d} \binom{n}{i} \leq \sum_{i=0}^{d} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$ multiply each term by sth ≥ 1

$\leq \sum_{i=0}^{n} \binom{n}{i} \left(\frac{n}{d}\right)^{d-i}$ add nonnegative terms to the sum

$= \left(\frac{n}{d}\right)^{d} \sum_{i=0}^{n} \binom{n}{i} \left(\frac{d}{n}\right)^{i}$ rearrange

$= \left(\frac{n}{d}\right)^{d} \left(1 + \frac{d}{n}\right)^{n}$ binomial theorem

$\leq \left(\frac{n}{d}\right)^{d} e^{d}$ e.g. from $1 - x \leq \exp(-x)$

(pause)

VC dimension of thresholds

$\mathcal{H} = \{h_a : a \in \mathbb{R}\}$

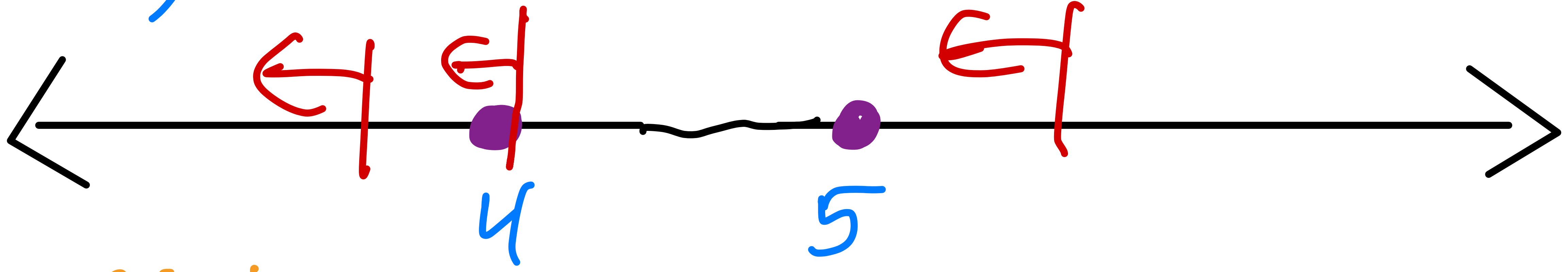
- Thresholds on $\mathbb{R}$: $h_a(x) = \mathbb{I}_{[x \le a]}$

$\exists\, C \subseteq \mathcal{X},\ |C| = d$, s.t. $\mathcal{H}$ shatters C

$\nexists\, C \subseteq \mathcal{X},\ |C| = d+1$ s.t. $\mathcal{H}$ shatters C

thresholds on $\mathbb{N}$

$\mathcal{H}' = \{h_z : z \in \mathbb{Z}\}$



$(0, 0) \leftarrow$ $h_3, h_2, h_{1.784}$

$(0, 1) \leftarrow$

$(1, 1) \leftarrow$

$\cancel{(1, 0)}$

$|\mathcal{H}_C|$

VCdim = 1

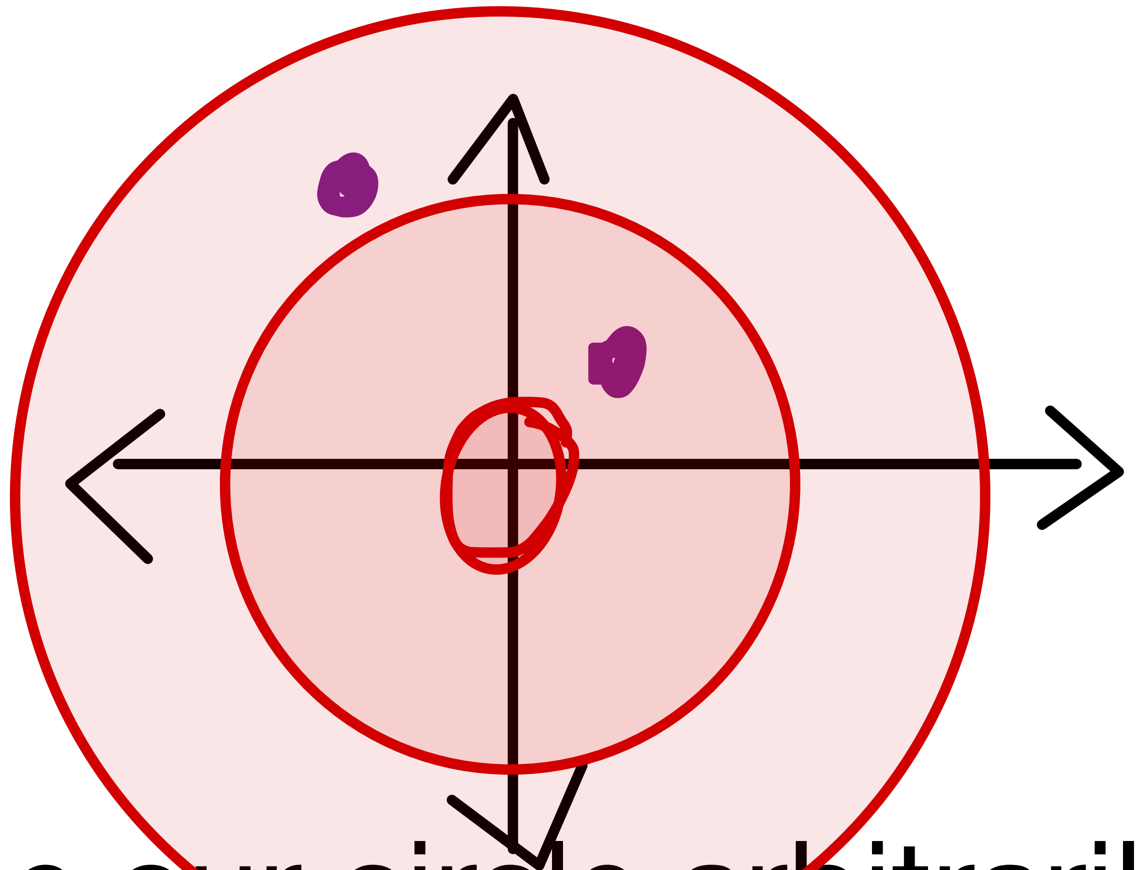
$\tau_{\mathcal{H}}(1) = 2$

$\tau_{\mathcal{H}}(2) = |\{(0,0), (0,1), (1,1)\}| = 3$

$\tau_{\mathcal{H}}(n) = n+1 \le \left(\frac{en}{\cancel{d}\ 1}\right)^{\cancel{d}\ 1} \le en$

VC dimension of circles

- From the homework: $h_r : \mathbb{R}^d \to \mathbb{R}$ given by $\mathbb{I}(\|x\| \leq r)$

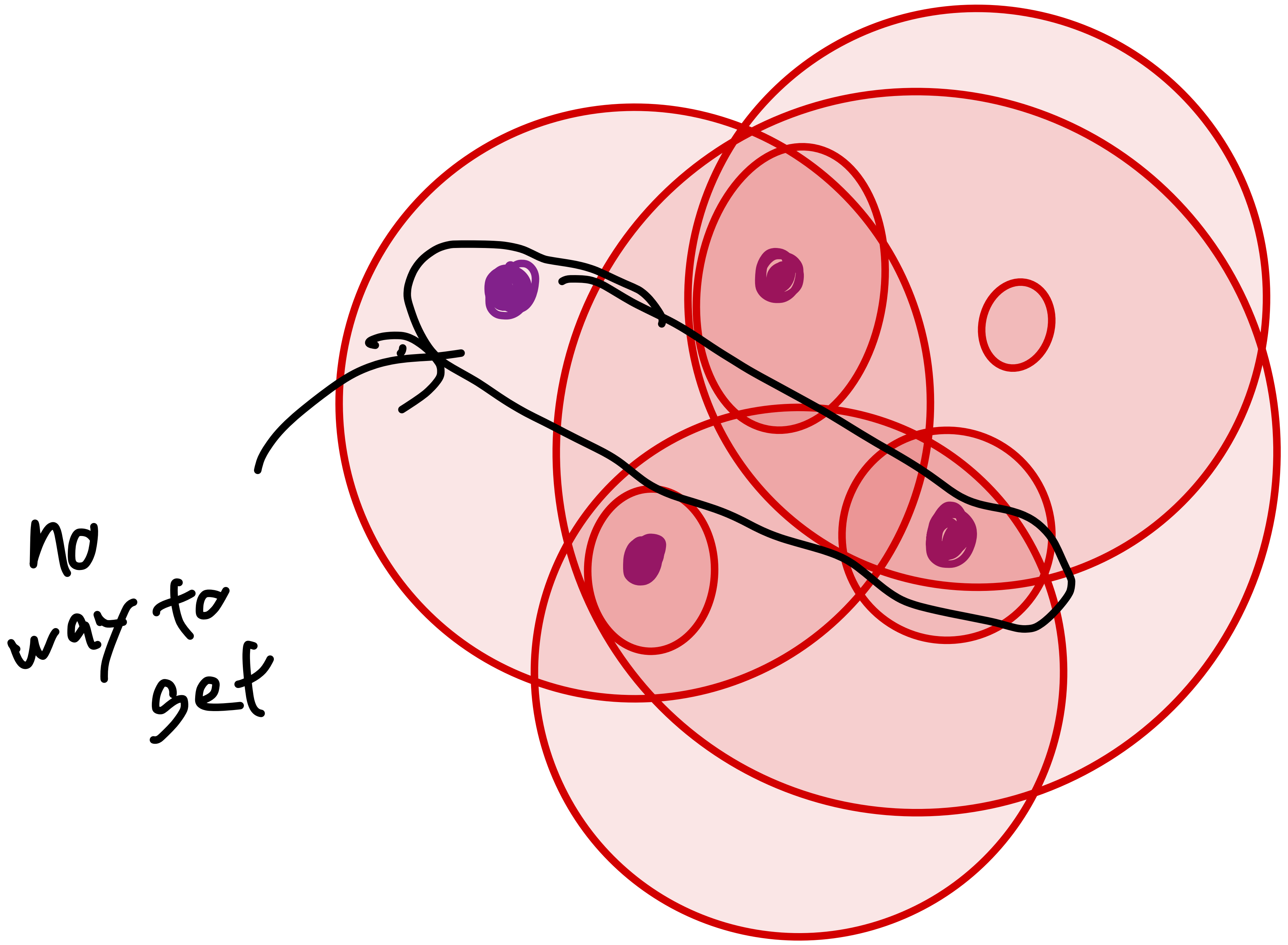


(0, 0)
(1, 0)
(1, 1)
$\cancel{(0, 1)}$

- What if we can place our circle arbitrarily? $h_{r,c} = \mathbb{I}(\|x - c\| \leq r)$

VC dimension of arbitrary circles

- What if we can place our circle arbitrarily? $h_{r,c} = \mathbb{I}(\|x - c\| \leq r)$ in $\mathbb{R}^2$



0, 0, 1
0, 1, 1
1, 0, 1
1, 1, 1

can shatter 3
can't shatter 4

VC dimension of finite classes

- Any $\mathcal{H}$ with $|\mathcal{H}| < \infty$

$$|\mathcal{H}_C| \leq |\mathcal{H}|$$

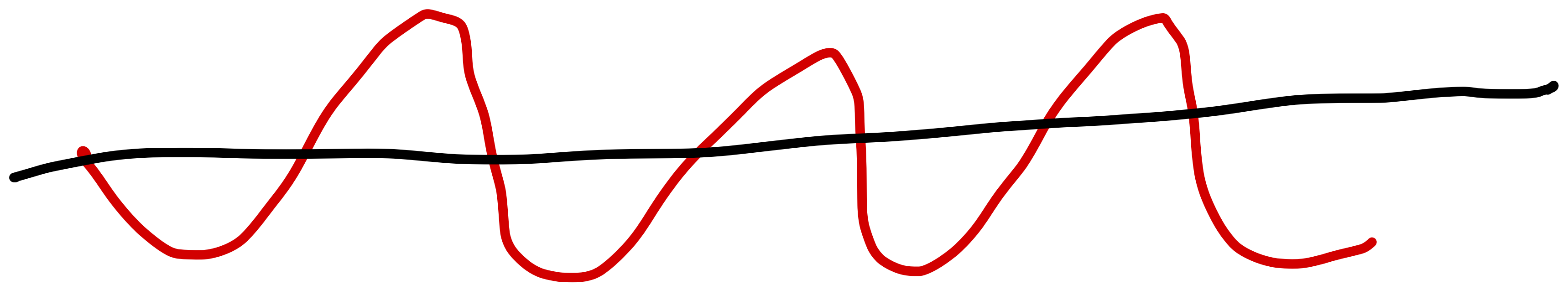
to shatter of size n, $|\mathcal{H}_C| = 2^n$

$$2^d \leq |\mathcal{H}_C| \leq |\mathcal{H}|$$

$$d \leq \log_2 |\mathcal{H}|$$

VC dimension of sign(sine)

- Let $h_w(x) = \mathbb{I}(\sin(wx) \geq 0)$ on $\mathbb{R}$



$$C = \{2^1, 2^2, \dots, 2^n\}$$
$$x_1, x_2, \dots, x_n$$

$$w = -\pi(0.y_1 y_2 y_3 y_4) = -\pi\left(\sum_{j=1}^{n} y_i 2^{-j}\right)$$

$$wx_i = -\pi\left(\sum_{j=1}^{n} y_j \underbrace{2^{i-j}}\right)$$

$$\sin(wx_i) = \sin\Big(\underbrace{-2k\pi} - \underbrace{\pi y_i} \underbrace{- \pi\sum_{j=i+1}^{n} y_j 2^{i-j}}_{<1}\Big)$$

$$\sin(-\pi - s(n)) \Rightarrow 0$$
$$\sin(0 - s(n)) \Rightarrow 1$$

$\therefore$ VC dim $= \infty$

Infinite VC dimension but can barely shatter anything

- Let $h_w(x) = \mathbb{I}(\sin(wx_1) \geq 0)$ on $\mathbb{R}^d$

$(0, \cdots)$

$(0, \cdots)$

$C_n = \{ (2^i, 0, \ldots, 0) : i \in [n] \}$

Recap

- The **growth function**, $\tau_{\mathscr{H}}(n) = \max_{C:|C|=n} |\mathscr{H}_C|$, bounds effective # of hyps

Recap

- The **growth function**, $\tau_{\mathscr{H}}(n) = \max\limits_{C:|C|=n} |\mathscr{H}_C|$, bounds effective # of hyps
- **Sauer-Shelah lemma**: when $n > d$, $\tau_{\mathscr{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$

Recap

- The **growth function**, $\tau_{\mathcal{H}}(n) = \max_{C:|C|=n} |\mathcal{H}_C|$, bounds effective # of hyps
- **Sauer-Shelah lemma**: when $n > d$, $\tau_{\mathcal{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$
- A generalization bound: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$

Recap

- The **growth function**, $\tau_{\mathcal{H}}(n) = \max_{C:|C|=n} |\mathcal{H}_C|$, bounds effective # of hyps
- **Sauer-Shelah lemma**: when $n > d$, $\tau_{\mathcal{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$
- A generalization bound: $\sup_{h\in\mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$
- Plugging together, get uniform convergence property for finite VCdim
  - and hence the "Fundamental Theorem of Learning"

Recap

- The **growth function**, $\tau_{\mathcal{H}}(n) = \max_{C:|C|=n} |\mathcal{H}_C|$, bounds effective # of hyps
- **Sauer-Shelah lemma**: when $n > d$, $\tau_{\mathcal{H}}(n) \leq (en/d)^d = \mathcal{O}\left(n^d\right)$
- A generalization bound: $\sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \leq \dfrac{4 + \sqrt{\log(\tau_{\mathcal{H}}(2n))}}{\delta\sqrt{2n}}$
- Plugging together, get uniform convergence property for finite VCdim
  - and hence the "Fundamental Theorem of Learning"
- Saw a bunch of VC calculations
  - Linear classifiers are d without intercept, $d+1$ with
  - Even stupid function classes can have infinite VC dim