Version Space (cont'd)

If $|\mathcal{H}|$ at most exponential then $\log |\mathcal{H}|$ polynomial and VS agent learns $\mathcal{H}$ online. This includes s-clause CNF's and s-term DNF's.

What about $\mathcal{H}$ covering *all* possible concepts on $O = \{ 0, 1 \}^n$, i.e.

$$\mathcal{C}(\mathcal{H}) = 2^O$$

We would not need to worry whether $C \in \mathcal{C}(\mathcal{H})$.

Since $|O| = 2^n$, we have $|\mathcal{H}| \geq 2^{|O|} = 2^{(2^n)}$, so $|\mathcal{H}|$ is super-exponential. So, nice try but no on-line learning.

Even some hypothesis classes which are more reasonable are super-exponential (we will see later).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

VC Dimension

Concept class $\mathcal{C}$ *shatters* $O' \subseteq O$ if any subset of O' coincides with $C \cap O'$ where $C \in \mathcal{C}$. A hypothesis class $\mathcal{H}$ shatters O' if $\mathcal{C}(\mathcal{H})$ shatters O'.

So O' is shattered by $\mathcal{C}$ (resp. $\mathcal{H}$) if its elements can be classified in all $2^{O'}$ possible ways by concepts from $\mathcal{C}$ (hypotheses from $\mathcal{H}$).

Vapnik-Chervonenkis Dimension

The *VC-dimension* of $\mathcal{C}$ (on O) denoted $\mathrm{VC}(\mathcal{C})$ is the cardinality of the largest subset of O shattered by $\mathcal{C}$. The VC-dimension of hypothesis class $\mathcal{H}$ is defined as $\mathrm{VC}(\mathcal{C}(\mathcal{H}))$, also denoted $\mathrm{VC}(\mathcal{H})$.

Note: the definition does not assume $\mathcal{C}$ or $\mathcal{H}$ finite.

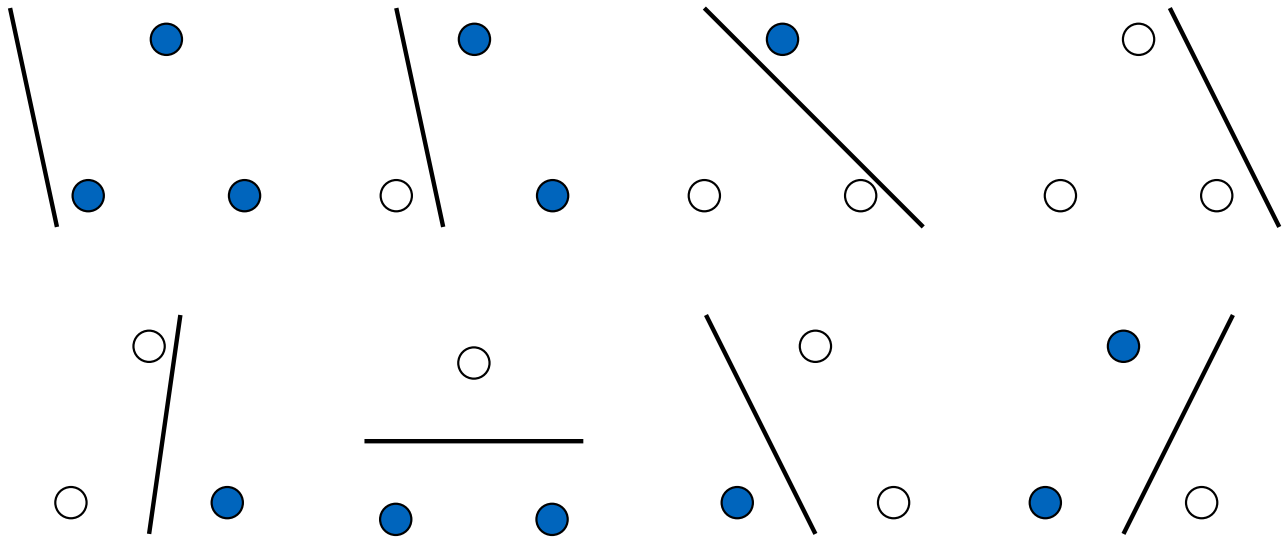
FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Determining VC-Dimension

- If *some* $O' \subseteq O$ shattered by $\mathcal{C}$ then $\mathrm{VC}(\mathcal{C}) \geq |O'|$.
- If *none* $O' \subseteq O$ shattered by $\mathcal{C}$ then $\mathrm{VC}(\mathcal{C}) < |O'|$.

Example: $\mathcal{C}$ = half-planes in $\mathbb{R}^2$ (i.e., linear separation)

- *Some* 3 points can be shattered



so $\mathrm{VC}(\mathcal{C}) \geq 3$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Determining VC-Dimension (cont'd)

- *No* 4 points can be shattered. Obvious if 3 in line. Otherwise two cases possible:

One point in the middle No point in the middle

In both cases, the shown subset cannot be separated by a line. So $\mathrm{VC}(\mathcal{C}) < 4$

We have $\mathrm{VC}(\mathcal{C}) \geq 3$ and $\mathrm{VC}(\mathcal{C}) < 4$, thus $\mathrm{VC}(\mathcal{C}) = 3$.

Lower Bounds on Mistake Bounds

No general lower bound on mistake *counts* as the agent may simply be lucky guessing right each time. But mistake *bounds* can be lower-bounded.

A mistake bound with no special assumptions cannot be lower than $|O|$ as each $o \in O$ may have an arbitrary class.

A mistake bound assuming only $C \in \mathcal{C} \subset 2^O$ for the target concept $\mathcal{C}$ cannot be smaller than $\mathrm{VC}(\mathcal{C})$ as there is a set $\left\{ o_1, o_2, \ldots, o_{\mathrm{VC}(\mathcal{C})} \right\} \subseteq O$ shattered by $\mathcal{C}$. So for the observation sequence $o_1, o_2, \ldots, o_{\mathrm{VC}(\mathcal{C})}$ and any sequence of agent's decisions $a_1, a_2, \ldots, a_{\mathrm{VC}(\mathcal{C})}$ there is a target concept $C \in \mathcal{C}$ by which all these decisions are wrong.

Corollary: an agent using hypothesis class $\mathcal{H}$ cannot be guaranteed to make fewer than $\mathrm{VC}(\mathcal{H})$ mistakes.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

I.I.D. Examples

So far we have maintained the general Markovian state transitions, so s_{k+1} was distributed according to

$$P_S(s_{k+1}|s_k, a_k)$$

The mistake bound model did not put any assumption on P_S. Now, we will assume that *s_{k+1} does not depend on s_k or a_k*, so all s_k are sampled from the same distribution

$$P_S(s)$$

so the s_k are *identically and independently distributed*, *i.i.d.* for short. As a consequence, observations o_k are also i.i.d from $P_O(o)$ where

$$P_O(o) = \sum_{s \in \{0,1\}} P_O(o|s) P_S(s)$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Hypothesis Accuracy and Error

The i.i.d. assumption lets us define the *accuracy of a hypothesis h* as the total probability of observations, which h decides correctly

$$\mathrm{acc}(h) = P_O(C \cap C(h))$$

and the *error of h* as

$$\mathrm{err}(h) = 1 - \mathrm{acc}(h)$$

In this setting, a large V^π is achieved when π uses a hypothesis with a small error, so the agent should minimize it. This gives rise to a new learning model.

Both acc and err are with respect to the distribution P_O and the target concept C, which we do not show explicitly in their notation.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

The PAC Learning Model

Informally, PAC-learning means finding a low-error hypothesis with high probability using a polynomial number of observations.

Probably Approximately Correct (PAC) Learning

Agent *probably approximately correctly (PAC) learns* $\mathcal{H}$ if for any $C \in \mathcal{C}(\mathcal{H})$ and numbers $0 < \epsilon, \delta < 1$, there is a $k < poly(n, 1/\epsilon, 1/\delta)$ such that the agent's hypothesis h_k has $\text{err}(h_k) \leq \epsilon$ with probability at least $1 - \delta$. $\mathcal{H}$ is *PAC-learnable* if there is an agent that PAC-learns it.

n is again the size of observations; with $O = \{ 0, 1 \}^n$, it is simply their arity.

Note that if O is finite and $\epsilon = \min_{o \in O} P_O(o)$, then with prob. at least $1 - \delta$, h_k is correct for *all* observations, i.e. $\text{err}(h_k) = 0$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Efficient and Proper PAC Learning

Analogically to the mistake-bound model, we say the agent PAC-learns $\mathcal{H}$ *efficiently* if it spends at most $poly(n, 1/\epsilon, 1/\delta)$ time between each percept and the subsequent action; if there is such an agent, $\mathcal{H}$ is *efficiently PAC-learnable*.

Note that in the definition of PAC learning we do not assume that $h_k \in \mathcal{H}$. In general, the agent may work with a hypothesis class larger than $\mathcal{H}$; for example with $\mathcal{H}$ =conjunctions, h_k may be a 3-CNF equivalent to the target conjunction. With the additional condition that $h_k \in \mathcal{H}$, we say that the agent *properly* (efficiently) PAC-learns $\mathcal{H}$, and if there is such an agent for $\mathcal{H}$, we say that $\mathcal{H}$ is *properly* (efficiently) PAC-learnable.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Generalizing Agent in the PAC Model

Let us analyze the generalizing agent again, this time with the i.i.d. assumption.

Let $O_l \subseteq O$ denote the set of all observations *inconsistent* with a literal l. We know already that a hypothesis h makes a mistake for an observation only if it has a literal inconsistent with it, so

$$\text{err}(h) \leq \sum_{l \in h} P_O(O_l)$$

With n variables, there are at most $2n$ literals in h so if $P_O(O_l) \leq \epsilon/2n$ for each literal $l \in h$ then $\text{err}(h) \leq \epsilon$. Call literal l *bad* if

$$P_O(O_l) > \frac{\epsilon}{2n} \tag{1}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Generalizing Agent in the PAC Model (cont'd)

Let l be bad. At time $k+1$, the probability that $l \in h_{k+1}$ is the probability that l is consistent with k i.i.d. observations (else it would have been removed), i.e. $(1 - P_O(O_l))^k$. Due to (1), we have

$$(1 - P_O(O_l))^k < \left(1 - \frac{\epsilon}{2n}\right)^k$$

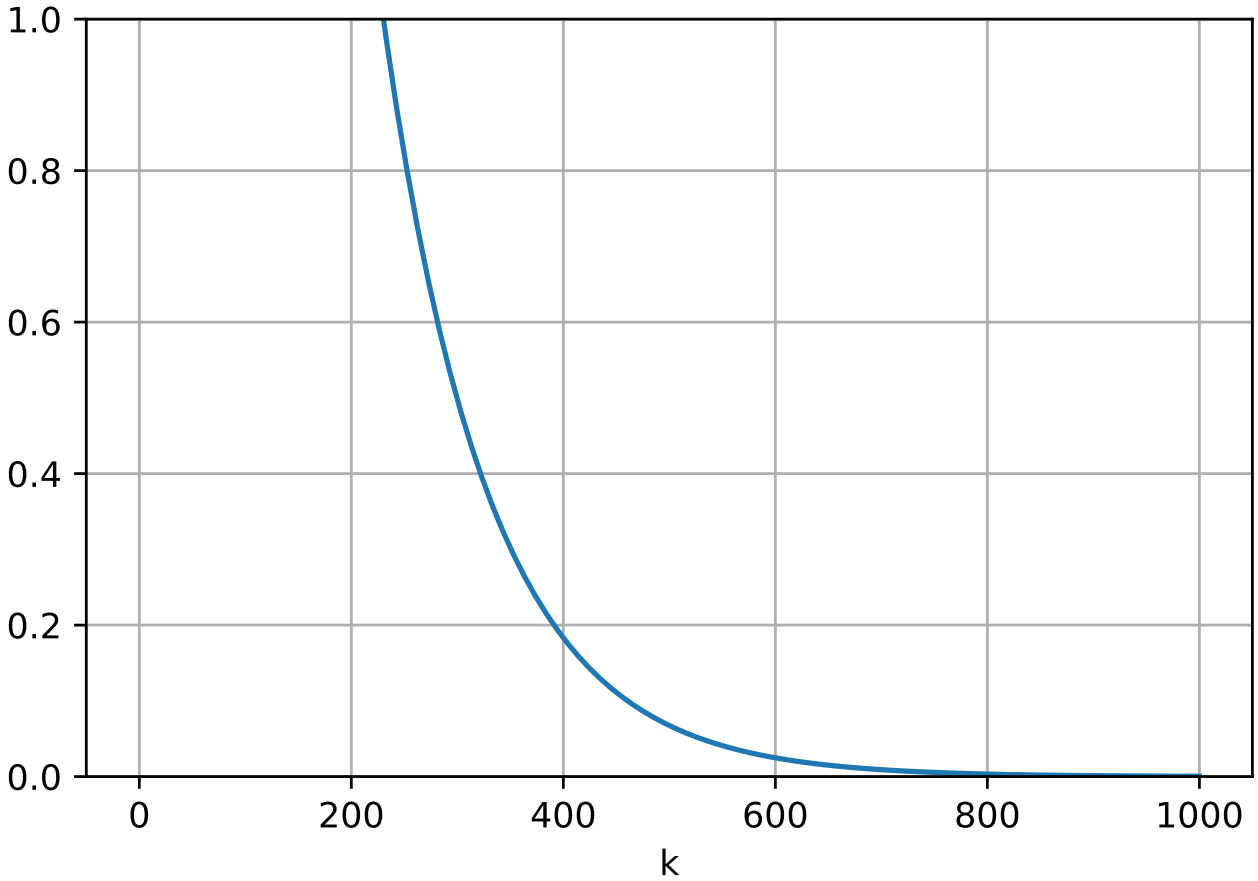
Prob. that *some* bad literal is consistent with the k observations is upper bounded by the above times the number of all literals so it is at most:

$$2n\left(1 - \frac{\epsilon}{2n}\right)^k < 2ne^{-k\frac{\epsilon}{2n}}$$

as there are at most $2n$ bad literals. For the right-hand side, we used $1 - x < e^{-x}$, $x \in (0; 1)$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

$2ne^{-k\frac{\epsilon}{2n}}$, $\epsilon = 0.1, n = 5$



Upper bound for probability that a bad literal remains in h_{k+1} of the generalizing agent with $n = 5$ variables. Note the looseness of this bound: from reasoning in the mistake-bound model, we know that $h_{2n+1} = h_{11}$ is already correct so it has no bad literal.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

PAC-Learning with the Generalization Agent

To satisfy PAC-learning conditions, we need to make it

$$2ne^{-k\frac{\epsilon}{2n}} < \delta$$

which is equivalent to

$$k \geq \frac{2n}{\epsilon} \ln \frac{2n}{\delta}$$

So for $k = \frac{2n}{\epsilon} \ln \frac{2n}{\delta}$, $\text{err}(h_{k+1}) \leq \epsilon$ with probability at least δ. Since $k + 1 \leq poly(n, 1/\delta, 1/\epsilon)$, *the agent PAC-learns conjunctions*.

It also learns them efficiently as it spends only $2n$ unit steps (checking each literal's consistency) on each observation.

Through adaptations we have discussed, it also efficiently PAC-learns disjunctions, s-CNF's and s-DNF's.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Standard On-line Agent

An on-line agent deciding by $\pi(h, o)$ is *standard* if it changes its hypothesis ($h_{k+1} \neq h_k$) if and only if h_k makes an error ($r_{k+1} = -1$). This includes the generalizing and separating agent but not e.g. version-space.

Consider any standard on-line agent with mistake bound M. Let $q \in \mathbb{N}$ and $k = Mq$. In the agent's sequence of hypotheses $h_1, h_2, \ldots h_{k+1}$, there must be a hypothesis h retained for at least q consecutive steps, i.e. $\exists K \leq k$ such that $h = h_K = h_{K+1} = \ldots = h_{K+q}$.

Assume for contradiction that all subsequences of identical hypotheses are shorter than q. Then there are more than M different hypotheses among $h_1, h_2, \ldots h_{k+1}$ because $k = Mq$. The agent changes its hypothesis only on a mistake, so this means it has made more than M mistakes up to time k. This is a contradiction because M is the error bound.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

PAC-Learning with any Standard On-Line Agent

The probability that a hypothesis h is consistent with q i.i.d. observations is $(1 - \text{err}(h))^q$. Call a hypothesis *bad* if $\text{err}(h) > \epsilon$. So a bad hypothesis is consistent with q i.i.d. observations with probability at most $(1 - \epsilon)^q$.

Consider any standard agent learning $\mathcal{H}$ online in the mistake-bound model, i.e., it makes at most $M < poly(n)$ mistakes. Let $k = Mq$. We already know that until time $k + 1$ the agent had a hypothesis h consistent with q consecutive observations.

This h is thus bad with probability at most $(1 - \epsilon)^q$. With $q = \frac{1}{\epsilon} \ln \frac{1}{\delta}$, this is $(1 - \epsilon)^q \leq e^{-q\epsilon} = e^{-\epsilon \frac{1}{\epsilon} \ln \frac{1}{\delta}} = \delta$. Since both M and q are $\leq poly(n, 1/\delta, 1/\epsilon)$, so is $k + 1 = Mq + 1$. Thus *if a standard agent (efficiently) learns $\mathcal{H}$ online it also (efficiently) PAC-learns it*.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Training Set and Batch Learning

Consider concept learning in an interaction with a *finite* horizon $m+1$. The agent's goal is to minimize the error of its last hypothesis $\mathrm{err}(h_{m+1})$.

Recall that in our interaction scenario, the class $s_k = c(o_k)$ of o_k is determined at $k+1$ as $|a_k + r_{k+1}|$. So at time step $m+1$ the agent has received exactly m observations with determined classes:

$$T = \{ \langle o_1, s_1 \rangle, \langle o_2, s_2 \rangle, \ldots, \langle o_m, s_m \rangle \} \tag{2}$$

which is called the *training (multi-)set*.

Rather than updating the hypothesis at each $k = 2, 3, \ldots m+1$, the agent can simply store the training set in its state (memory), and only at time $m+1$ compute a hypothesis from T. We will call this *batch learning*.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Consistency with the Training Set

The hypothesis a PAC-learning agent computes from any training set is consistent with it.

Let h be the hypothesis computed from T (2) and assume for contradiction that h misclassifies some $o \in \{ o_1, \ldots, o_m \}$. P_O and $0 < \delta < 1$ can be arbitrary so set them such that $\delta < \prod_{k=1}^{m} P_O(o_k)$, i.e. T is received with probability greater than δ. This implies that $P_O(o_k) > 0$ for $1 \leq k \leq m$, so also $P_O(o) > 0$. Because of that and since $0 < \epsilon < 1$ can be arbitrary, we can set ϵ such that $\epsilon < P_O(o)$. Since h misclassifies o, it follows $\mathrm{err}(h) \geq P_O(o)$ and because $P_O(o) > \epsilon$, we have $\mathrm{err}(h) > \epsilon$. This happens when T is received, i.e. with probability greater than δ. This contradicts PAC-learning conditions.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Consistent Agent

Consider a general *consistent agent* learning $\mathcal{H}'$ and equipped with some hypothesis class $\mathcal{H}$. Given a training set T of size m, it produces an arbitrary $h \in \mathcal{H}$ consistent with T.

Note that producing a consistent hypothesis for any given T is only possible if the target concept C is in $\mathcal{C}(\mathcal{H})$. $C \in \mathcal{C}(\mathcal{H})$ is arbitrary, so

$$\mathcal{C}(\mathcal{H}) \supseteq \mathcal{C}(\mathcal{H}') \tag{3}$$

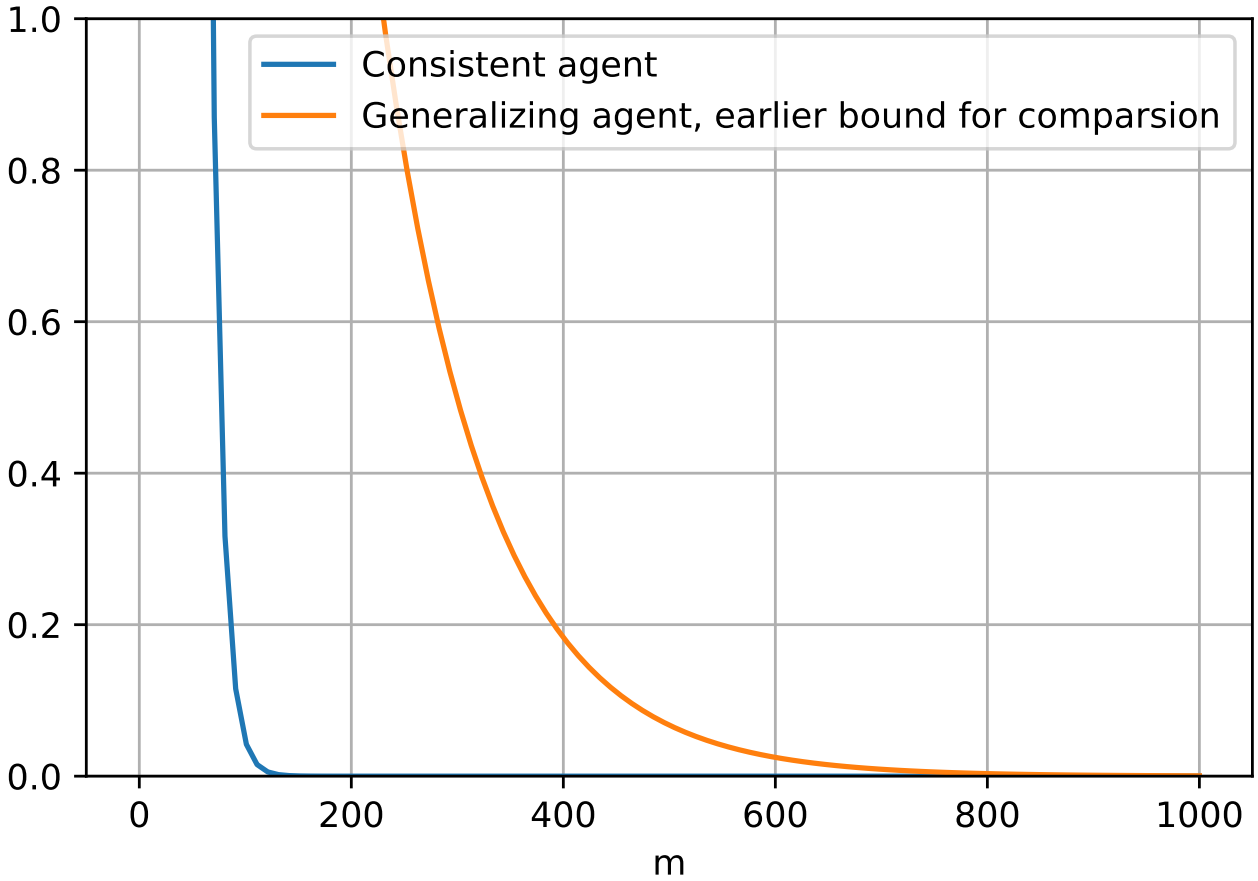
is a necessary condition.

The probability that a hypothesis $h \in \mathcal{H}$ consistent with T is bad ($\mathrm{err}(h) > \epsilon$) is $(1 - \mathrm{err}(h))^m < (1 - \epsilon)^m$. There are at most $|\mathcal{H}|$ hypotheses so the probability that *some* bad hypothesis is consistent with T is at most

$$|\mathcal{H}|(1 - \epsilon)^m < |\mathcal{H}|e^{-\epsilon m}$$

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

$|\mathcal{H}|e^{-\epsilon m}$, $\epsilon = 0.1$, $n = 5$



Upper bound on prob. that err(h) $> \epsilon$ with a consistent agent and $\mathcal{H}$ = conjunctions (blue). The bound applies also to the generalizing agent, which is consistent, and is tighter than the previous bound derived specifically for it (orange).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

PAC-Learning with any Consistent Agent

$|\mathcal{H}|e^{-\epsilon m}$ is smaller than δ if

$$m \geq \frac{1}{\epsilon} \ln \frac{|\mathcal{H}|}{\delta}$$

Since m is polynomial in $1/\epsilon$ and $1/\delta$, *PAC-learns* $\mathcal{H}$ if m is further polynomial in n. The only factor on the right-hand side depending on n is $\ln |\mathcal{H}|$, so the condition is

$$\ln |\mathcal{H}| \leq poly(n)$$

i.e., $|\mathcal{H}|$ is at most *exponential* in n.

To learn $\mathcal{H}'$ *properly* we further require $\mathcal{H} \subseteq \mathcal{H}'$, which together with (3) implies $\mathcal{H}' = \mathcal{H}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

Sizes of Some $\mathcal{H}$ (You do the math)

$\mathcal{H}$	$\lvert\mathcal{H}\rvert$ in increasing size	
s-conjunctions, s-disjunctions	$\mathcal{O}\left[(2n)^s\right]$ (incl. self-resolv.)	poly
s-depth decision trees	$2(2n)^{2^s-1}$	poly
conjunctions, disjunctions	2^{2n} (3^n if no self-resolving)	exp
s-term DNF, s-clause CNF	$\mathcal{O}\left[(3^n)^s\right]$	exp
s-CNF, s-DNF	$\mathcal{O}\left[2^{\binom{2n}{s}}\right] = \mathcal{O}\left[2^{(n^s)}\right]$	exp
dec. trees, DNF, CNF, ...	$\geq 2^{(2^n)}$ (# all concepts)	super-exp

Notes:

s-term DNF, s-clause CNF. At most s non-self-resolving terms (clauses) of unlimited size.

DNF, CNF. $|\mathcal{H}| = 2^{(3^n)}$. There are only $2^{(2^n)}$ possible concepts on $O = \{0,1\}^n$, so $\mathcal{H}$ has equivalent pairs in it.

decision trees. Can express any concept so $|\mathcal{H}| \geq 2^{(2^n)}$

s-depth decision trees. See next slide.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE