# Version Space

## Theorem 1

*A finite hypothesis class $\mathcal{H}$ is learnable online from $X$ if $\lg|\mathcal{H}| \leq poly(n_X)$.*

Proof: the **version space** algorithm below has the mistake bound $\lg|\mathcal{H}|$ so if $\lg|\mathcal{H}| \leq poly(n_X)$ the version space learns $\mathcal{H}$ online from $X$.

Define a decision policy for a *set of hypotheses* as the *majority vote* among the hypotheses in it

$$\mathcal{H}(x) = \begin{cases} 1 \text{ if } |\{\, h \in \mathcal{H} \;:\; h(x) = 1 \,\}| > |\mathcal{H}|/2 \\ 0 \text{ otherwise} \end{cases} \tag{1}$$

Instead of a single hypothesis $h_k$, the version-space agent maintains a set $\mathcal{H}_k$ of hypotheses at each time $k \in \mathbb{N}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Version Space (cont'd)

At $\forall k \in \mathbb{N}$

$$y_{k+1} = \mathcal{H}_k(x_k)$$

where $\mathcal{H}_1 = \mathcal{H}$ and $\mathcal{H}_{k+1}$ keeps exactly those hypotheses from $\mathcal{H}_k$ which gave the correct decision for $x_k$:

$$\mathcal{H}_{k+1} = \{\, h \in \mathcal{H}_k : h(x_k) = \overline{y}_k \,\} \tag{2}$$

where $\overline{y}_k = |y_k + r_{k+1}|$ is the true class according to the target hypothesis.

At least half of the hypotheses from $\mathcal{H}_k$ is deleted by (2) on each mistake ($r_k = -1$), because at least half of them were wrong according to (1). In the worst case, the single remaining hypothesis is the target hypothesis, which will no longer make mistakes. So $\lg |\mathcal{H}_1| = \lg |\mathcal{H}|$ is the maximum number of mistakes.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Online Learnability due to Polynomial $\lg |\mathcal{H}|$

Theorem (1) implies online learnability of the hypothesis classes

$$\mathcal{H}_1 \subset \mathcal{H}_2 \subset \mathcal{H}_3$$

$\mathcal{H}_1$= monotone conjunctions (clauses, respectively): $|\mathcal{H}| = 2^n$ (each atom present or absent in it).

$\mathcal{H}_2$= contingent conjunctions (clauses, respectively): $|\mathcal{H}| = 3^n$ (each atom absent, present in a positive literal or present in a negative literal)

$\mathcal{H}_3$= conjunctions (clauses, respectively): $|\mathcal{H}| = 2^{2n}$ (each of $2n$ literals present or absent)

from $X = \{0,1\}^n$ when $n_X = n$ due to $\lg |\mathcal{H}| \leq poly(n)$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Version Space vs. Algorithms for Specific $\mathcal{H}$

We already know these classes are *efficiently* learnable online from $X = \{0,1\}^n$ by Winnow with mistake bound $\mathcal{O}(\lg n)$. Version space has bound $\lg |\mathcal{H}| = \mathcal{O}(n)$ in all these cases and is *not efficient* as it loops over an exponential-size set $\mathcal{H}$ in (1) and (2).

Contingent conjunctions (clauses) are also *efficiently* learnable online from contingent conjunctions (clauses) by generalization with bound $\mathcal{O}(n)$ where $n_X = n$ is the maximum number of variables in the conjunctions (clauses). The version space bound $\lg |\mathcal{H}| = \mathcal{O}(n)$ is the same but again, version space is *not effificent*.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Online Learnability due to Polynomial $\lg|\mathcal{H}|$ (cont'd)

Further classes learnable online from $X = \{0,1\}^n$:

$\mathcal{H} =$ s-conjunctions or s-clauses: not only $\lg|\mathcal{H}| \leq poly(n)$ but $|\mathcal{H}| \leq poly(n)$. So version space learns them online with logarithmic mistake bound and *efficiently* because (1) and (2) can be evaluated in $poly(n)$-time.

$\mathcal{H} =$ s-DNF or s-CNF: These include a subset of the $poly(n)$-number of s-conjunctions (s-clauses, respectively), so $|\mathcal{H}| = 2^{poly(n)}$, i.e. $\lg|\mathcal{H}| = poly(n)$. So version space learns them online with logarithmic mistake bound but not efficiently. In contrast, they can be learned online efficiently by reduction to monotone disjunctions (conjunctions).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# VC Dimension

We say that concept class $\mathcal{C}$ **shatters** a set of observations $X' \subseteq X$ if for every subset $X'' \subseteq X'$ there is a concept $C \in \mathcal{C}$ such that $C \cap X' = X''$.

In other words, $X'$ is shattered by $\mathcal{C}$ if it can be split by concepts from $\mathcal{C}$ in all $2^{|X'|}$ possible ways.

## Vapnik-Chervonenkis Dimension

The **VC-dimension of concept class $\mathcal{C}$ on $X$** denoted $\mathrm{VC}(\mathcal{C})$ is

$$\max\{\ |X'|\ :\ \mathcal{C} \text{ shatters } X',\ X' \subseteq X\ \}$$

For a hypothesis class $\mathcal{H}$, we abbreviate $\mathrm{VC}(\mathcal{C}(\mathcal{H}))$ as $\mathrm{VC}(\mathcal{H})$ and call the latter the VC-dimension of $\mathcal{H}$.

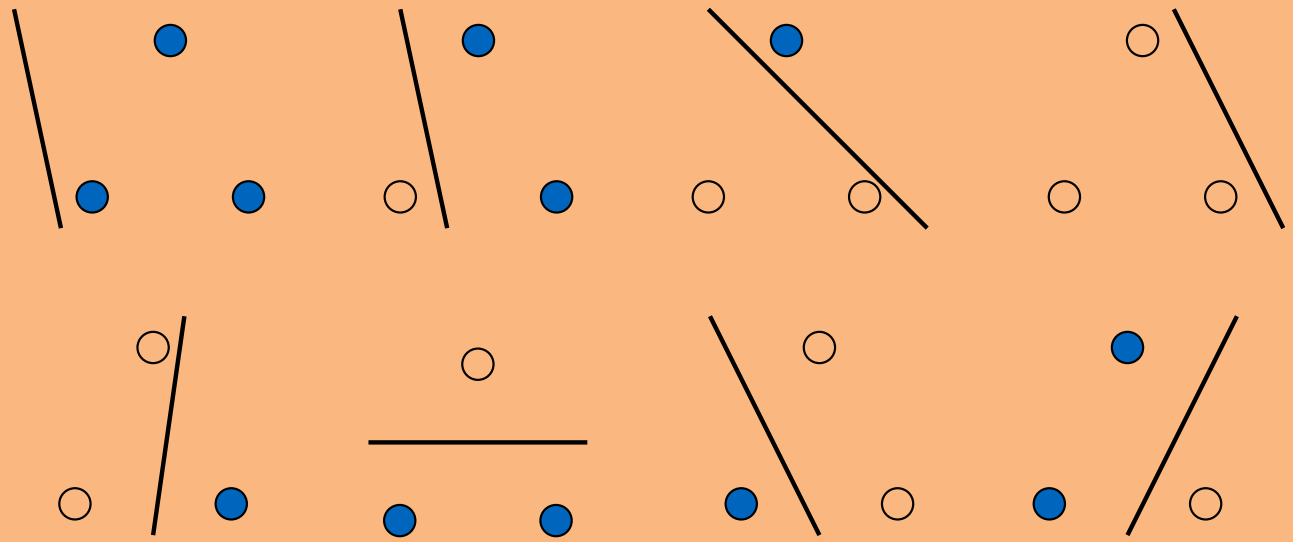
FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Example: Determining VC-Dimension

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

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 colored 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 $\mathrm{VC}(\mathcal{C}) = 3$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Poly VC-Dimension Necessary for Learnability

**Theorem 2**

*Concept class $\mathcal{C}$ on $X$ is learnable online from $X$ only if $\mathrm{VC}(\mathcal{C}) \leq poly(n_X)$.*

Proof: There exists a set of $\mathrm{VC}(\mathcal{C})$ observations from $X$ shattered by $\mathcal{C}$ so there exists a sequence $x_{\leq \mathrm{VC}(\mathcal{C})}$ of observations such that for any sequence of the learner's decisions $y_2, y_3, \ldots, y_{\leq \mathrm{VC}(\mathcal{C})+1}$ there is a target concept $C \in \mathcal{C}$ on which this sequence results in $\mathrm{VC}(\mathcal{C})$ mistakes, i.e., $r_2, r_3, \ldots, r_{\mathrm{VC}(\mathcal{C})+1} = \underbrace{(-1, -1, \ldots -1)}_{\mathrm{VC}(\mathcal{C})}$, thus $\sum_{k=1}^{\infty} |r_k| \geq \mathrm{VC}(\mathcal{C})$. So if $\mathcal{C}$ is learnable, i.e. if $\sum_{k=1}^{\infty} |r_k|$ is polynomially bounded, $\mathrm{VC}(\mathcal{C})$ is also polynomially bounded.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# $\lg|\mathcal{H}|$ vs. $\mathrm{VC}(\mathcal{H})$

Theorem (2) holds for any concept class $\mathcal{C}$, including a concept class $\mathcal{C}(\mathcal{H})$ induced by a hypothesis class $\mathcal{H}$. So if $\mathcal{H}$ is learnable online from $X$ then $\mathrm{VC}(\mathcal{H}) \leq poly(n_X)$.

Combining this with Theorem (1), we get that $\mathrm{VC}(\mathcal{H})$ must be polynomial whenever $\lg|\mathcal{H}|$ is.

Note that in Theorem (1), we cannot replace "hypothesis class" with "concept class" and $\mathcal{H}$ with $\mathcal{C}$, because there may be concepts in $\mathcal{C}$ that have no finite description by a hypothesis, thus the assumption in the Theorem would not be sufficient for online learnability.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Standard Agent

An agent is called **standard** if it changes its hypothesis iff an error is made, i.e. $h_{k+1} \neq h_k$ iff $r_{k+1} = -1$. So Winnow and the generalization agent are both standard while version space is not.

## Lemma 1

*Let $r_{\leq K}$ be a reward sequence of a standard agent and $h_{\leq K}$ be its sequence of hypotheses. Denote $M = \sum_{k=1}^{K} |r_k|$. There is a hypothesis $h$ retained for at least $\frac{K}{M+1}$ consecutive steps in $h_{\leq K}$, i.e.*

$$h_{\leq K} = h_1, h_2, \ldots \underbrace{h, h, \ldots h}_{\text{at least } \frac{K}{M+1} \text{ times}}, \ldots h_K$$

Proof: exercise

*If $\frac{K}{M+1}$ is non-integer than "at least $\frac{K}{M+1}$" means the same as "at least the nearest integer largest than $\frac{K}{M+1}$".*

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# I.I.D. Observations and Hypothesis Error

Assume (??) for the rest of the concept-learning chapter.

Given target concept $\overline{C}$, define the **error of hypothesis** $h$ as

$$\text{err}(h) = P(\overline{C} \triangle C(h))$$

where $\triangle$ is the symmetric set difference. *Recall the definition of C(h).*

So err($h$) is the

- total probability of the error region on $X$ with respect to the distribution (??).
- probability that policy $h(x)$ is not the true class of $x$, i.e., not the optimal action (??).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# The PAC Learning Model

Unlike the mistake bound model, the PAC-learning model does not require a finite bound on mistakes. It requires finding a *low-error hypothesis* with *high probability* using a *polynomial number of observations*.

## Probably Approximately Correct (PAC) Learning Model

In the concept classification protocol, an agent **probably approximately correctly (PAC) learns $\mathcal{C}$ from $X$** if for any target concept from $\mathcal{C}$ on $X$, an arbitrary distribution (??) and arbitrary numbers $0 < \epsilon, \delta < 1$, there is a $K < poly(n_X, 1/\epsilon, 1/\delta)$ such that with probability at least $1 - \delta$, the agent's hypothesis $h_K$ has $\text{err}(h_K) \leq \epsilon$. It PAC-learns $\mathcal{C}$ from $X$ **efficiently**, if in addition, the time taken to compute an action from an observation is also at most polynomial in $n_X, 1/\epsilon, 1/\delta$.

$\mathcal{C}$ is **PAC-learnable** if there is an algorithm that PAC-learns it.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Online Learnability Implies PAC-Learnability

Call a hypothesis $h$ **consistent** with $x \in X$ if $h(x)$ is the true class of $x$, i.e. $h(x) = 1$ iff $x$ is in the target concept.

## Theorem 3

*If $\mathcal{C}$ is learnable online from $X$ by a standard agent, then $\mathcal{C}$ is PAC-learnable from $X$.*

*The standard-agent assumption is not needed in the theorem but it is a rather weak assumption making the proof much easier.*

Proof: Online learnability of $\mathcal{C}$ from $X$ means $M = \sum_{k=1}^{\infty} |r_k| \leq poly(n_X)$. Due to Lemma (1), for an arbitrary $K \in \mathbb{N}$, some hypothesis $h$ was retained by the agent for $q = \frac{K}{M+1}$ steps before time $K$. Since the learning agent is standard, $h$ is consistent with $q$ consecutive observations.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Online Learnability Implies PAC-Learnability (cont'd)

To show that the agent PAC-learns $\mathcal{C}$, we want to show that $\text{err}(h) \leq \epsilon$ with probability at least $1 - \delta$ when the observations are i.i.d. from (??).

Call $h$ **bad** if $\text{err}(h) > \epsilon$. The probability that a $h$ is consistent with an observation from (??) is $1 - \text{err}(h)$, so if $h$ is bad, it is at most $(1 - \epsilon)$. Because $h$ is consistent with $q$ such i.i.d. consecutive observations, the probability that $h$ is bad is at most $(1 - \epsilon)^q$.

*Note that it is important for the above reasoning that h is consistent with q observations, which are consecutive. For example, we could not 'pick' an arbitrary selection of q correctly classified observation from an arbitrarily long history $x_{\leq k}$ (that would not be a series of Bernoulli trials).*

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Online Learnability Implies PAC-Learnability (cont'd)

We want the probability $(1-\epsilon)^q$ to be less than $\delta$. Use the upper bound

$$(1-\epsilon)^q < e^{-\epsilon q} \tag{3}$$

which holds for all $\epsilon > 0$. So $(1-\epsilon)^q < \delta$ if $e^{-\epsilon q} \leq \delta$. The latter is satisfied if

$$q = \frac{1}{\epsilon} \ln \frac{1}{\delta} \tag{4}$$

because $e^{-\epsilon \frac{1}{\epsilon} \ln \frac{1}{\delta}} = \delta$. Since $q = \frac{K}{M+1}$, we have $K = (M+1)q$ where $M+1 \leq poly(n_X)$ and $q \leq poly(\frac{1}{\epsilon}, \frac{1}{\delta})$, therefore $K \leq poly(n_X, \frac{1}{\epsilon}, \frac{1}{\delta})$.

Thus with probability *at least* $\delta$, $h_K$ is *not bad* ($\text{err}(h) \leq \epsilon$) for $K \leq poly(n_X, \frac{1}{\epsilon}, \frac{1}{\delta})$ so the agent PAC-learns $\mathcal{C}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# PAC-Learnability Implies Consistent Learning

The following theorem states that in the definition of PAC-Learning, the hypothesis $h_K$ is necessarily consistent with all observations up to $K-1$.

## Theorem 4

*Let an agent PAC-learn $\mathcal{C}$ from $X$. Then for any target concept from $\mathcal{C}$ on $X$, an arbitrary distribution (??) and arbitrary numbers $0 < \epsilon, \delta < 1$ and $K \in \mathbb{N}$, the condition $\mathrm{err}(h_K) \leq \epsilon$ with probability at least $1-\delta$ implies that $h_K$ is consistent with all observations in $x_{<K}$.*

Proof: exercise

# Consistent Agent

**$\mathcal{H}$-consistent learning agent** is an agent whose $h_{k+1}$ ($k \in \mathbb{N}$) is an arbitrary hypothesis from $\mathcal{H}$ that is consistent with $x_{\leq k}$ if such a hypothesis exists.

When the target concept is chosen from a concept class $\mathcal{C}$, the condition

$$\mathcal{C} \subseteq \mathcal{C}(\mathcal{H}) \tag{5}$$

(where $\mathcal{C}(\mathcal{H})$ is defined by (??)) guarantees that a hypothesis consistent with any $x_{\leq k}$ exists in $\mathcal{H}$.

Note that an $\mathcal{H}$-consistent agent may make mistakes even when (5) is satisfied but after each mistake ($r_{k+1} = -1$), $h_{k+1}$ is chosen such that it is consistent with all of $x_{\leq k}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Consistent Agent (cont'd)

*Recall we are assuming observations $x_k$, $k \in \mathbb{N}$ to be i.i.d.!*

## Lemma 2

*For the hypothesis $h_{k+1}$ of a $\mathcal{H}$-consistent agent satisfying* (5), $\mathrm{err}(h_{k+1}) \leq \epsilon$ *with probability at least* $1 - \delta$ *if*

$$k \geq \frac{1}{\epsilon} \ln \frac{|\mathcal{H}|}{\delta} \tag{6}$$

Proof: Due to the assumption (5), $\mathcal{H}$ contains a hypothesis consistent with an arbitrary sequence of observations $x_{\leq k}$, so the agent can pick one at each $k+1$ and the agent has at most $|\mathcal{H}|$ possible choices. The probability that one such chosen hypothesis $h_{k+1} \in \mathcal{H}$ consistent with $x_{\leq k}$ is bad ($\mathrm{err}(h_{k+1}) > \epsilon$) is $(1 - \mathrm{err}(h_{k+1}))^k < (1 - \epsilon)^k$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Consistent Learning with with Polynomial lg $|\mathcal{H}|$

The probability that *some* of the $|\mathcal{H}|$ possible choices is bad is thus at most

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

where we used the bound (3). This is smaller than $\delta$ if (6) holds. This completes the proof. We get the following theorem as a corollary.

## Theorem 5

*An $\mathcal{H}$-consistent learning agent satisfying (5) PAC-learns from $X$ any concept class $\mathcal{C}$ on $X$ if* $\lg |\mathcal{H}| \leq poly(n_X)$.

Proof: By Lemma (2), an $\mathcal{H}$-consistent learning agent satisfying (5) has $\mathrm{err}(h_k) \leq \epsilon$ with probability at least $1-\delta$ for $k$ polynomial in $\frac{1}{\epsilon}, \frac{1}{\delta}$ (6). The requirement $\lg |\mathcal{H}| \leq poly(n_X)$ means that $k$ is also polynomial in $poly(n_X)$, thus the agent PAC-learns $\mathcal{C}$.

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Consistent Learning with with Polynomial VC($\mathcal{H}$)

## Lemma 3

*For the hypothesis $h_{k+1}$ of a $\mathcal{H}$-consistent agent satisfying* (5), $\mathrm{err}(h_{k+1}) \leq \epsilon$ *with probability at least* $1 - \delta$ *if*

$$k \geq \max\left\{\frac{4}{\epsilon}\lg\frac{2}{\delta}, \frac{8 \cdot \mathrm{VC}(\mathcal{H})}{\epsilon}\lg\frac{13}{\epsilon}\right\} \tag{7}$$

Proof omitted due to length and technicality.

*$\mathrm{VC}(\mathcal{H})$ may be finite and polynomial even when $|\mathcal{H}|$ is infinite and $\lg|\mathcal{H}|$ is not defined. Even for a finite $|\mathcal{H}|$,* (7) *may give a better (smaller) bound than* (6). *(Exercise problem)*

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE

# Consistent Learning with with Polynomial VC($\mathcal{H}$) (cont'd)

A corollary of Lemma (3) is analogical to Theorem (5).

**Theorem 6**

*An $\mathcal{H}$-consistent learning agent satisfying (5) PAC-learns from $X$ any concept class $\mathcal{C}$ on $X$ if $\mathrm{VC}|\mathcal{H}| \leq poly(n_X)$.*

The proof is also analogical to that of Theorem (5), except it relies on Lemma (3) instead of Lemma (2).

FACULTY OF ELECTRICAL ENGINEERING CTU IN PRAGUE