Announcements

- Final 7-8:15 PM, Wed. 12/15 here
- Q/A session 11-noon Mon. 12/13 2405SC
- Projects (for 4 credits) due Tue. 12/7
  - Code
  - Sample I/O (if it doesn't work, say so)
  - Paper discussing
    - What you did & why
    - What you learned
    - How you would do it differently given…

Computational Learning Theory How Much Data is Enough?

- Training set is evidence for which $h \in H$ is
  - Correct: [Simple, Proper, Realizable??] learning
  - Best: *Agnostic* learning
- Remember: training set = labeled independent samples from an underlying population
- Suppose we perform well on the training set
- How well will perform on the underlying population?
- This is the *test accuracy* or *utility* of a concept (not how well it classifies the training set)

What Makes a Learning Problem Hard?

- How do we measure “hard”?
- Computation time?
- Space complexity?
- What is the valuable resource?
- **Training examples**
- Hard learning problems require more training examples
- Hardest learning problems require the entire example space to be labeled

[Simple] Learning

- PAC formulation
- Probably Approximately Correct
- Example space X sampled with a fixed but unknown distribution $\mathcal{D}$
- Some target concept $h^* \in H$ is used to label an iid (according to $\mathcal{D}$) sample S of N examples
- Finite H
- Algorithm: return any $h \in H$ that agrees with all N training examples **S** |**S**| = N
- Choose N sufficiently large that with high confidence ($1-\delta$) h has accuracy of at least $1-\varepsilon$ $0 < \varepsilon, \delta \ll 1$

$$N \geq \frac{1}{\varepsilon}\left(\ln\frac{1}{\delta} + \ln|H|\right)$$

Simple Learning
(simple derivation)

- What is the probability that a bad hypothesis looks good? (need to bound this to be $\leq \delta$)
  - Bad h: true error of h > $\boldsymbol{\varepsilon}$
  - Looks good: correct on our training set of N examples
- Hypothesis h, h* $\in$ **H** and x $\in$ **X** drawn with $\mathcal{D}$
  - h is bad: $Pr_{\mathcal{D}}(h(x) \neq h^*(x)) > \varepsilon$
  - h looks good on **S**: $\forall s \in \mathbf{S}$ $h(s) = h^*(s)$ $|\mathbf{S}| = N$
- What is
  - Probability of bad h getting a single $x \sim \mathbf{X}_{\mathcal{D}}$ correct?
  - $Pr_{\mathcal{D}}(h(x) = h^*(x)) \leq 1-\varepsilon$
  - Probability of two $x \sim \mathbf{X}_{\mathcal{D}}$ correct?
  - $Pr_{\mathcal{D}}(h(x) = h^*(x)) \leq (1-\varepsilon)^2$
  - Probability of N $x \sim \mathbf{X}_{\mathcal{D}}$ correct?
  - $Pr_{\mathcal{D}}(h(x) = h^*(x)) \leq (1-\varepsilon)^N$

Simple Learning
(simple derivation)

- Probability of N $x \sim \mathbf{X}_{\mathcal{D}}$ correct from bad h is
  $Pr_{\mathcal{D}}(h(x) = h^*(x)) \leq (1-\varepsilon)^N$
- This bounds prob. of a single bad h masquerading as good on N – not enough; too weak…
- We must limit that *ANY* $h \in \mathbf{H}$ tricks us
- These probabilities can be no worse than exclusive
  union bound (very useful): $Pr(A \vee B) \leq Pr(A) + Pr(B)$
- Prob. that any bad $h \in \mathbf{H}$ masquerades as good is less than…
  $|H| (1-\varepsilon)^N$ (can't be any more than |H| bad hypotheses…)
- We want to be at least $1 - \delta$ confident that this does *not* happen
- It is sufficient that $|H| (1-\varepsilon)^N \leq \delta$

(the rest is just math…)
[solve for N – one more little trick…]

Simple Learning
(simple derivation)

- It is sufficient that

$$|H|\,(1-\varepsilon)^N \leq \delta$$

Or $\ln |H| + N \cdot \ln (1-\varepsilon) \leq \ln \delta$

- Recall $e^{-y} > 1-y$ (for $y > 0$) so $\ln (1-\varepsilon) < -\varepsilon$ and substituting gives a safer δ
- It suffices that $\ln |H| - N \cdot \varepsilon \leq \ln \delta$

Or $N \geq (\ln \delta - \ln |H|)/ -\varepsilon$

Or $N \geq (1/\varepsilon)\,(-\ln \delta + \ln |H|)$

Or $N \geq (1/\varepsilon)\,(\ln (1/\delta) + \ln |H|)$ (very loose bound)

See Text section 18.5

Agnostic Learning

- Same thing but no guarantee $h^* \in H$
- Possibly, no h is consistent over S
- With confidence at least 1-δ,
  find an h that is no more than ε worse
  than the best $h \in H$.
- Bernoulli events: h's error rate on S (|S|=N);
  (like repeated coin flips, Pr(h wrong) is coin weighting)
  relate sample error rate to the true error rate
- Chernoff bound for a sequence of N Bernoulli events:

$$P(\mu_S > \mu_D + \varepsilon) \leq e^{-2N\varepsilon^2}$$

- This bounds the probability that the sample accuracy of an arbitrary h evaluated on S is very misleading:

$$P\left(error_S(h) > error_D(h) + \varepsilon\right) \leq e^{-2N\varepsilon^2}$$

- We can again bound the probability that *any* one has a misleading error:

$$P\left((\exists h \in H) error_S(h) > error_D(h) + \varepsilon\right) \leq |H| e^{-2N\varepsilon^2}$$

- We need this to be bounded by δ:

$$P\left((\exists h \in H) error_S(h) > error_D(h) + \varepsilon\right) \leq |H| e^{-2N\varepsilon^2} \leq \delta$$

- Solving for N:

$$N \geq \frac{1}{2\varepsilon^2}\left(\ln|H| + \ln\frac{1}{\delta}\right)$$

Intuition for Why it Works

"Choose N sufficiently large that with confidence of at least (1-δ), h has an accuracy of at least (1-ε)."

- In some regions of X we don't care how well h performs
- h need be close only where it matters
- $|S| = N \Rightarrow D_S$ approximates $\mathcal{D}$ such that:
  - Where D_S is uncertain, $Pr_{\mathcal{D}}(x)$ is low
  - Where $Pr_{\mathcal{D}}(x)$ is high, D_S approximates $\mathcal{D}$ well

What about Infinite H?

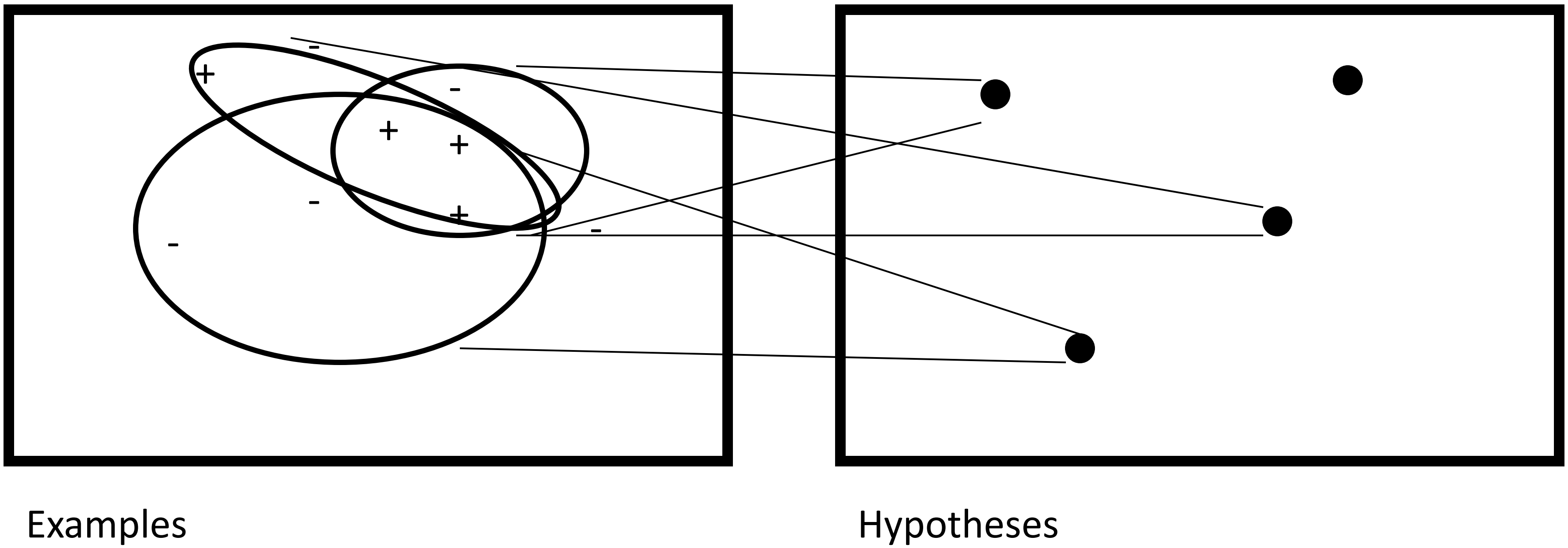
- Essentially, VC(H) plays the role of ln|H|
- For learning w/ finite H:

$$N \geq \frac{1}{\varepsilon}\left(\ln\frac{1}{\delta} + \ln|H|\right)$$

- For learning w/ infinite H:

$$N \geq \frac{c}{\varepsilon}\left(\ln\frac{1}{\delta} + VC(H)\cdot\ln\frac{1}{\varepsilon}\right)$$

Hypotheses as Partitioning Functions $h_i : X \rightarrow \{+,-\}$



Given a set of n labeled examples, is there a hypothesis consistent with it?

Suppose we change the labels – is there still a consistent hypothesis?

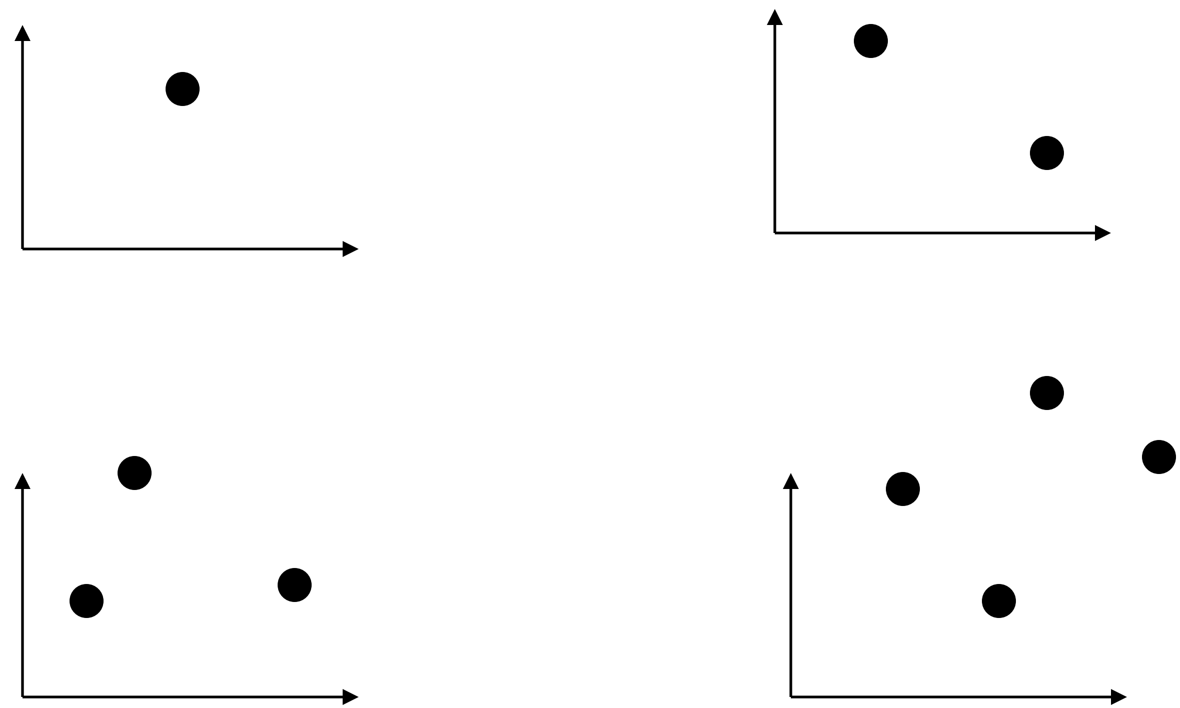
What is the largest n for which the answer is "yes" ?

This is the Vapnik-Chervonenkis dimension of the hypothesis space VC(H)

Capacity & VC Dimension

- VC is most common but there are other measures of *capacity*
- VC(H) is the cardinality of the largest set of examples *shattered* by H
- An example set is shattered by a hypothesis set iff every classification labeling assignment of the examples, is consistent with some element of H

2d Perceptron VC Dimension



Thus the VC dimension of a 2-d perceptron is 3

The largest set of points that can be labeled arbitrarily

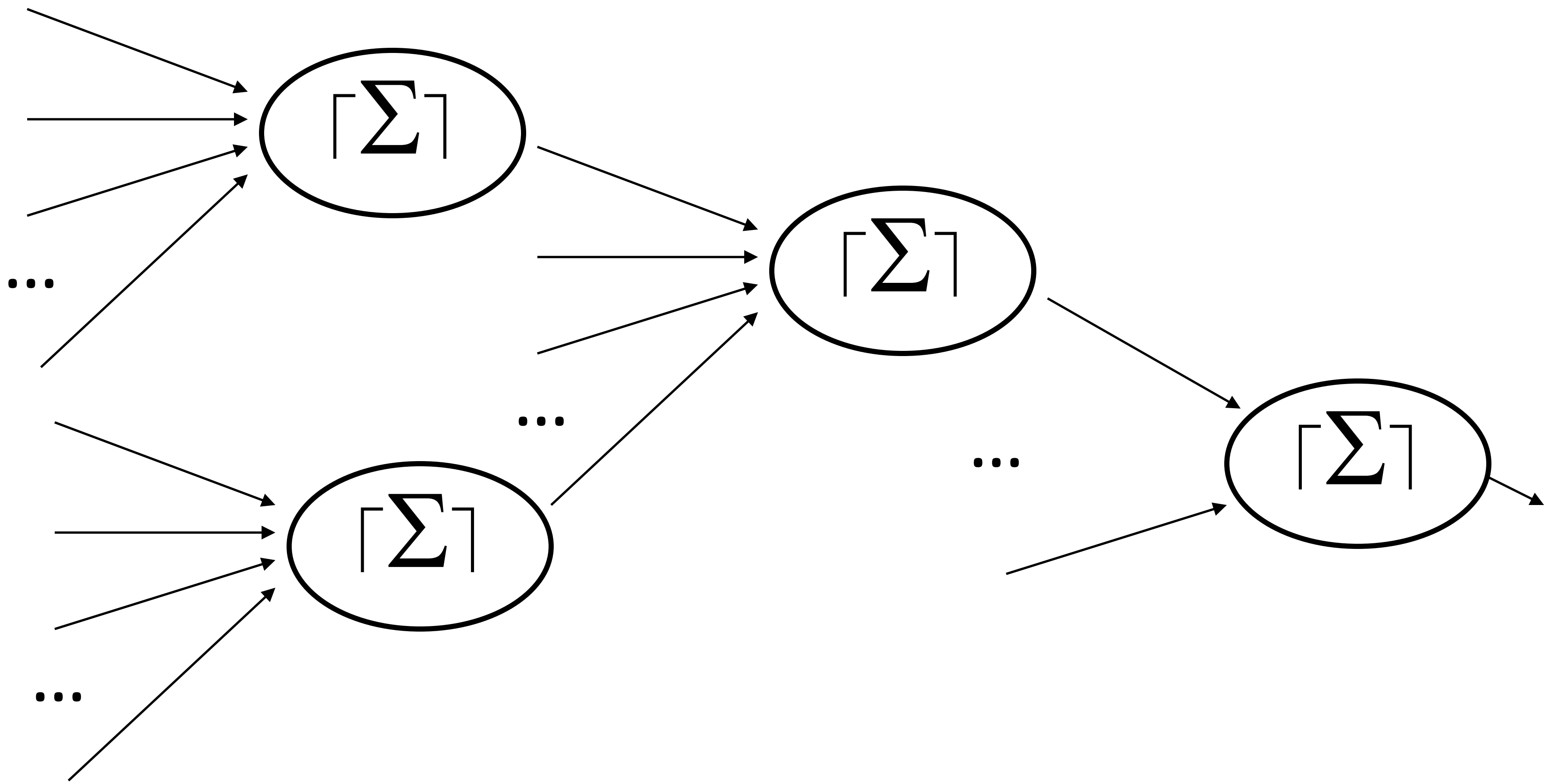
Note infinite |H| but low expressiveness

Examples of VC Dimensions

- Intervals on the real line
- 2
- Linear half-spaces in the plane
- 3
- d dimensional hyperplane
- d+1
- Axis-aligned rectangles in the plane
- 4
- Feed forward artificial neural net
- $O(v \cdot s \cdot \log(s))$ s units; v is VC of component

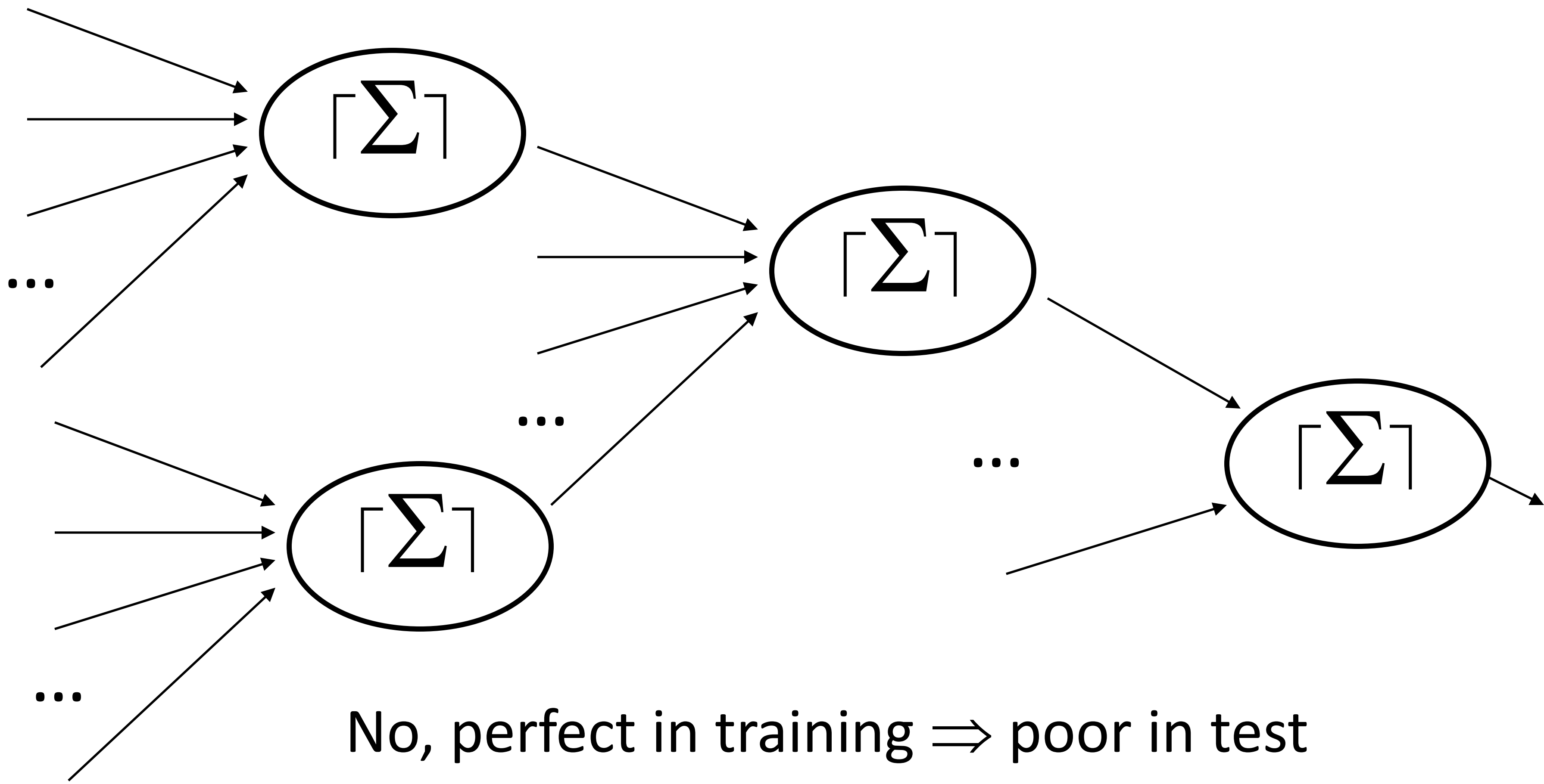
With enough units, an ANN (MLP) can learn any assignment of labels

(is this a good thing?)



With enough units, an ANN (MLP) can learn any assignment of training labels

Is this a good thing?



No, perfect in training $\Rightarrow$ poor in test

VC Dimension of a Concept Class

- Can be challenging to prove
- Can be non-intuitive
- Signum(sin($\omega \cdot x$)) on the real line
- Convex polygons in the plane

Learnability

- Often the hypothesis space (or concept class) is syntactically parameterized

    n-Conjuncts, k-DNF, k-CNF, m of n, MLP w/ k units,…

- The concept class is *PAC learnable* if there exists an algorithm whose running time grows no faster than polynomially in the natural complexity parameters: $1/\varepsilon$, $1/\delta$, others
- Clearly, polynomially-bounded growth in the minimum number of training examples is a necessary condition.