What is learnable?

Readings for today

- Fulop, S. A., & Chater, N. (2013). Learnability theory. Wiley Interdisciplinary Reviews: Cognitive Science, 4(3), 299-306.

Topics

1. The two parts of a statistical model

2. Understanding what is learnable

The two parts of a statistical model

Fundamental form of a statistic

Fundamental form of a statistic

relational function

$$Y = f(X)$$

output input

Sample

$$S = \{(X_i, Y_i)\}_{i=1}^{n}$$

- h : hypothesis → prediction based on X
- g : learner → system that finds the best h for a particular S

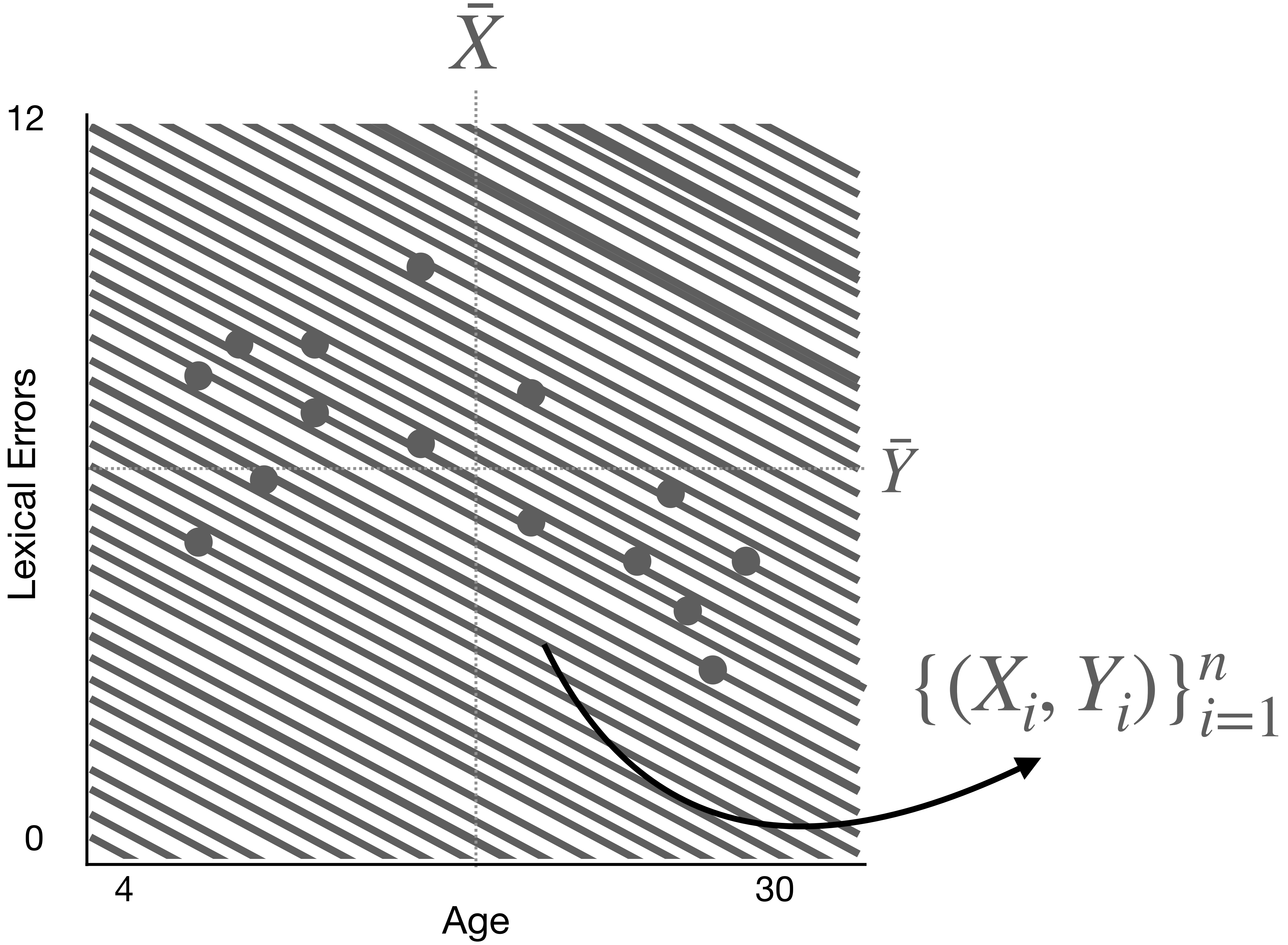
The learning problem: $h \,\&\, g \to f$

<u>Hypothesis</u>

$$h : Y = \hat{\beta}_1 X + \hat{\beta}_0$$

<u>Learner</u>

$$g : \beta = (X'X)^{-1}X'Y$$



Empirical Risk Minimization

Risk: how well a hypothesis does in practice with new (test) data.

$$R(h) = \underbrace{\ell(h(X), Y)}_{\text{loss function}}$$

Goal of learner

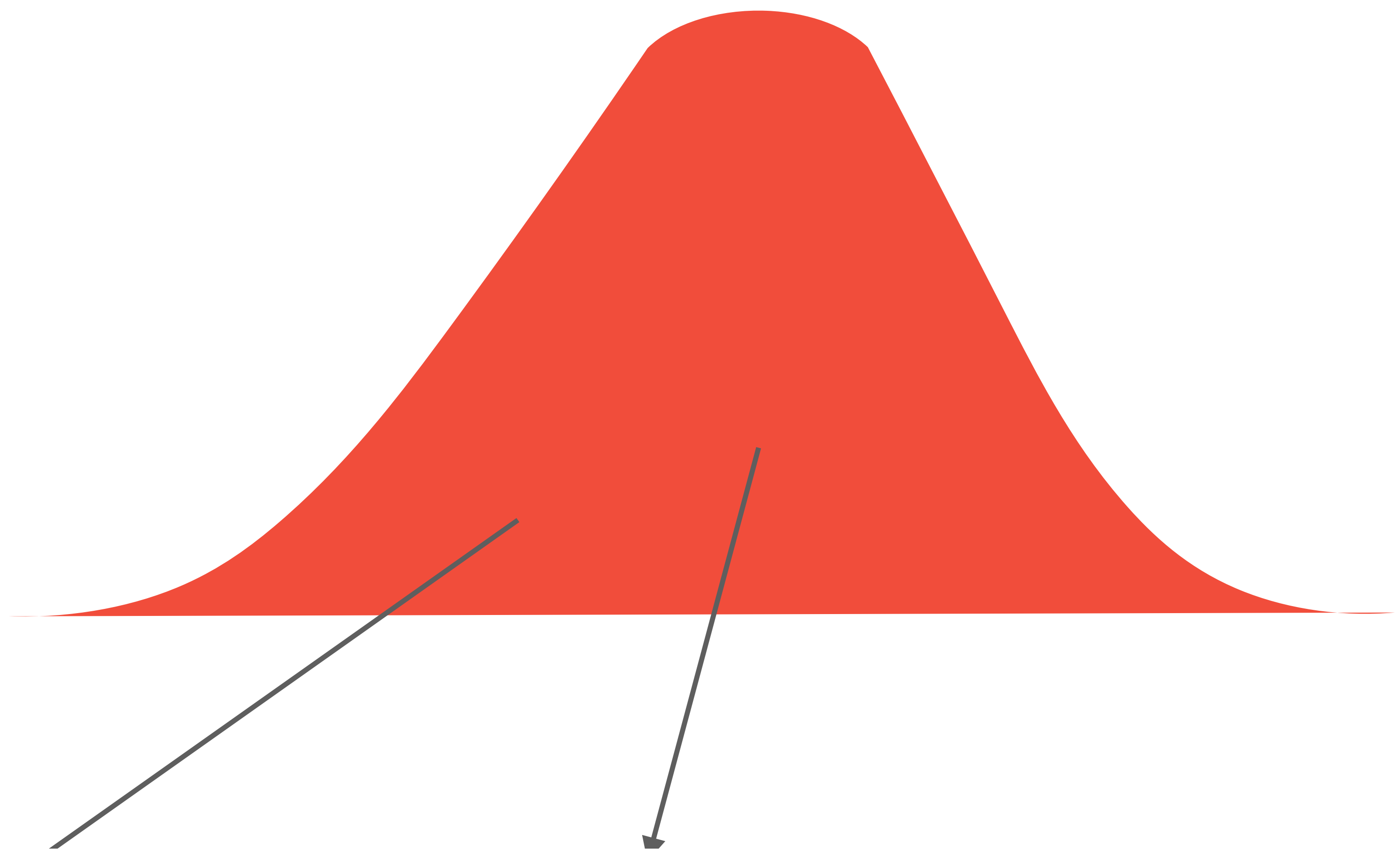
$$g = \underbrace{\arg\min}_{h \in H} R(h)$$

$$\ell(h(X), Y) = \begin{cases} 1, & h(X) \neq Y \\ 0, & h(X) = Y \end{cases}$$

Empirical Risk Minimization

Expected Risk

$$E_{\text{risk}}\left(h, n, P\right) = \int_{\underbrace{(\mathbf{X},\mathbf{Y})_n}_{\text{train}}} \underbrace{R(h)}_{\text{risk}} \underbrace{dP_{(X,Y)_n}}_{\text{distribution}}$$



Assumption: Both the training and test data come from the same distribution.

Understanding what is learnable

What is learnable?

Learnability theory: Can the true h be learned?

Determine whether the true h is able to be learned given a particular data scenario (e.g., particular $P_{(\mathbf{X},\mathbf{Y})}$)?

Computational theory: Can the solution for the true h be computed?

P: Problem can be *computed* in polynomial time. → *We can only work with "P hard" problems*

NP: Problem requires non-deterministic polynomial time.

(Fulop & Chater 2013)

Bayesian Learning

Bayes Rule:

$$P(H_i \,|\, \underbrace{\{(X,Y)\}_{i=1}^{n}}_{\text{D}}) = \frac{P(D\,|\,H_i) \quad P(H_i)}{P(D)}$$

Naive prior:

$$P(H_i) = U(-\infty, +\infty)$$

$$P(H_i\,|\,D) = P(D\,|\,H_i)$$

H_i: individual hypothesis

(Fulop & Chater 2013)

Bernstein-von Mises Theorem

If:
- (X, Y) are independent and identically distributed
- limited to a finite space of outcomes.

Assuming:
- The true h is learnable at all.

Then:
- The posterior probability converges to the true h among almost all samples.

e.g., testing & training sets

(Fulop & Chater 2013)

Model selection

$$h_{true} \rightarrow h_{best}$$

<u>Generalization:</u> Given a fixed training data set, find the model that *best* predicts future unseen (test) data set.

$$E_{\text{risk}}(h, n, P) = \underbrace{\int_{(\mathbf{X},\mathbf{Y})_n}}_{\text{train}} \underbrace{\int_{(\mathbf{X},\mathbf{Y})}}_{\text{test}} \underbrace{\ell(h(X), Y)}_{\text{loss function}} \underbrace{dP_{X,Y}}_{\text{distribution}} \underbrace{dP_{(X,Y)_n}}_{\text{distribution}}$$

(Fulop & Chater 2013)

Model selection

<u>Bayesian evidence:</u>

Relative evidence in favor of H_m over H_j

$$\frac{P(H_m | D)}{P(H_j | D)} = \frac{P(H_m)}{P(H_j)} \times \frac{P(D | H_m)}{P(D | H_j)}$$

<u>Naive prior:</u>

$$P(H_i) = U(-\infty, +\infty)$$

<u>Neyman-Pearson Lemma:</u>

When probabilities are known, the optimal decision rule for selecting one H over another is a likelihood ratio test.

(Fulop & Chater 2013)

Probably Approximately Correct (PAC) Learning

Q: How do you get "good enough" learning so as to be useful?

PAC learning requires a learner to:

1. Approximate the true h
2. Be computationally feasible (P problem)

Approximately: A hypothesis $h \in H$ is approximately correct if its error over the training data $P_{(X,Y)}$ is bounded by some ϵ, with $0 \leq \epsilon \leq \frac{1}{2}$.

Probably: The h is probably approximately correct at a generalization error rate of δ, if its prediction accuracy is $1 - \delta$, with $0 \leq \delta \leq \frac{1}{2}$.

(Valiant 1984; Fulop & Chater 2013)

Probably Approximately Correct (PAC) Learning

<u>Bounding the sample size:</u>

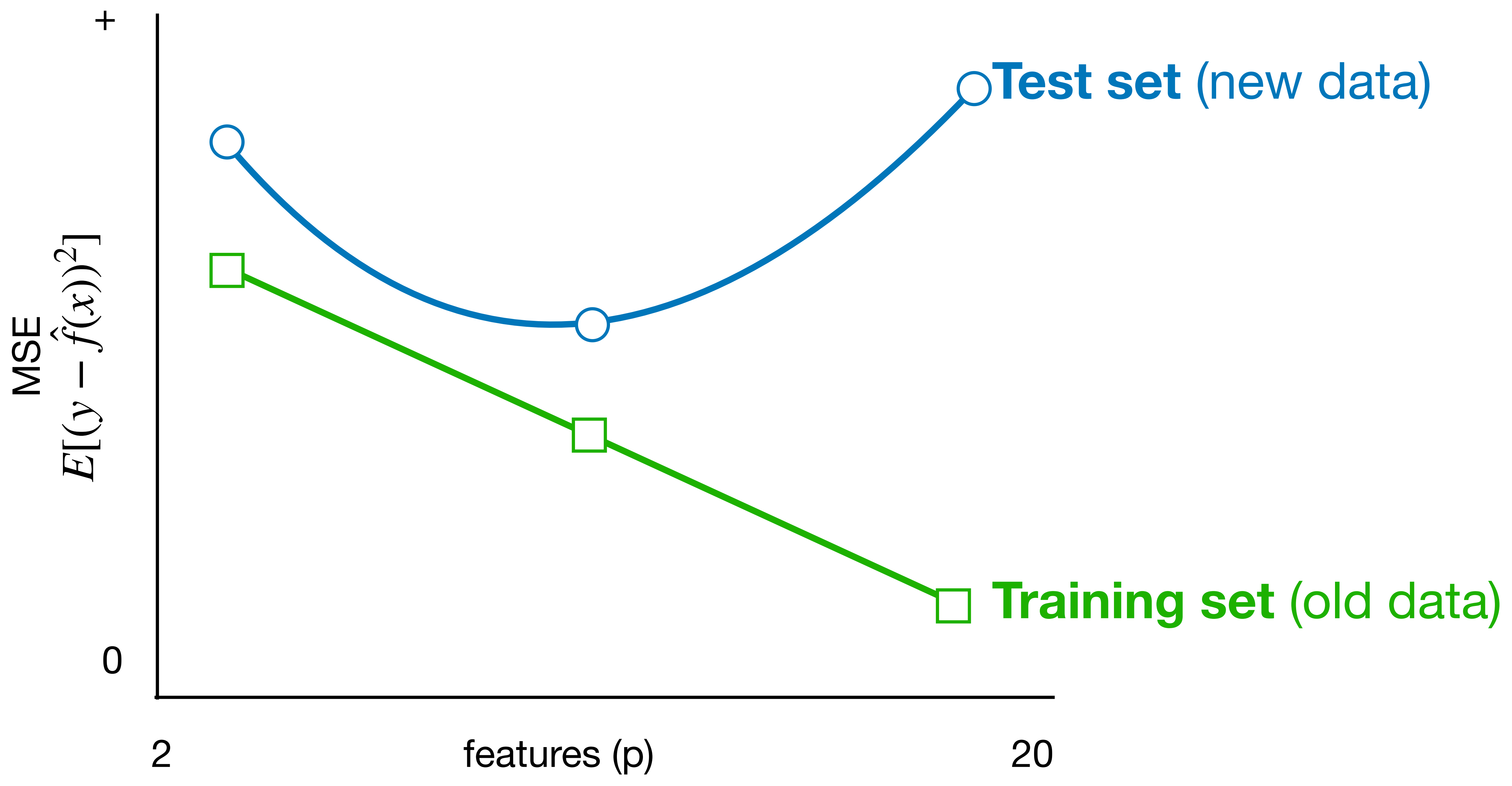
$$n \geq \frac{1}{\epsilon}(\ln|H| + \ln\frac{1}{\delta})$$

e.g., complexity of model

- as $\downarrow \epsilon$ & $\downarrow \delta$, or $\uparrow H$, then $\uparrow n$

<u>The bias-variance tradeoff:</u>

- $\delta \rightarrow$ Test error
- $\epsilon \rightarrow$ Training error



(Valiant 1984;(Fulop & Chater 2013)

Take home message

- A statistical model consists of two parts: 1) a hypothesis, h, that predicts an output from an input and 2) a learner, g, that finds the best h for a given sample.
- Before running any analysis, you first must confirm whether your h (aka- model) is learnable and computable given the expected data set it will be applied to.