Learning Theory: Why ML Works

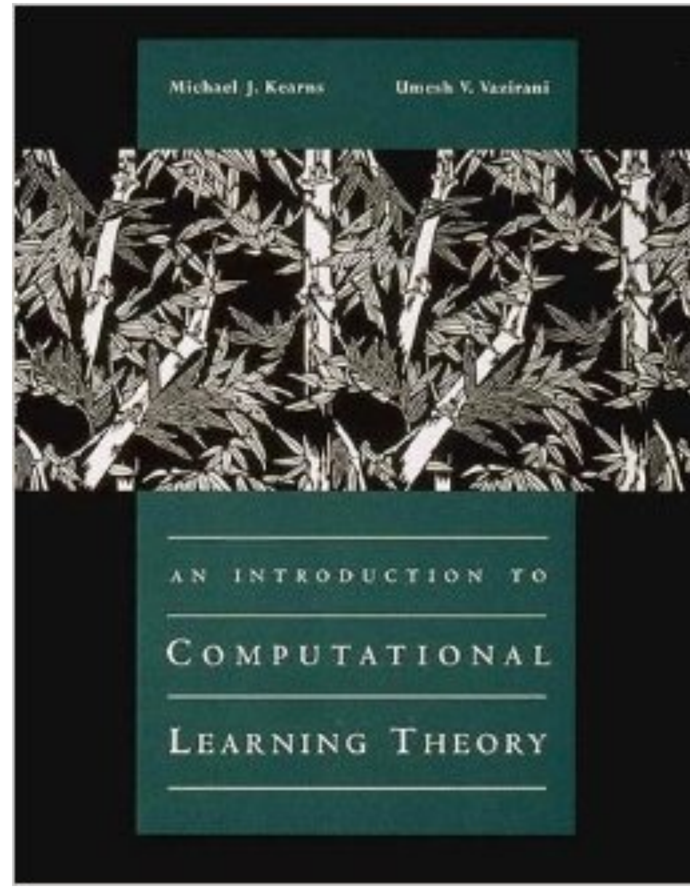
These slides were assembled by Byron Boots, with only minor modifications from Eric Eaton's slides and grateful acknowledgement to the many others who made their course materials freely available online. Feel free to reuse or adapt these slides for your own academic purposes, provided that you include proper attribution.

Computational Learning Theory

Entire subfield devoted to the mathematical analysis of machine learning algorithms

Has led to several practical methods:

- PAC (probably approximately correct) learning → boosting
- VC (Vapnik–Chervonenkis) theory → support vector machines



Annual conference: Conference on Learning Theory (COLT)

Computational Learning Theory

Fundamental Question: What general laws constrain a system's ability to learn?

Seeks theory to relate:

- Probability of successful learning
- Number of training examples
- Complexity of hypothesis space
- Accuracy to which target function is approximated
- Manner in which training examples should be presented

Based on slide by Tom Mitchell

Sample Complexity

Assume that $f : \mathcal{X} \mapsto \{0, 1\}$ is the target function

How many training examples are sufficient to learn the target function f ?

1. If learner proposed instances as queries to teacher
   - Learner proposes instance x, teacher provides f(x)
2. If teacher (who knows f) provides training examples
   - Teacher provides labeled examples in form <x, f(x)>
3. If some random process (e.g., nature) proposes instances
   - Instance x generated randomly, teacher provides f(x)

Based on slide by Tom Mitchell

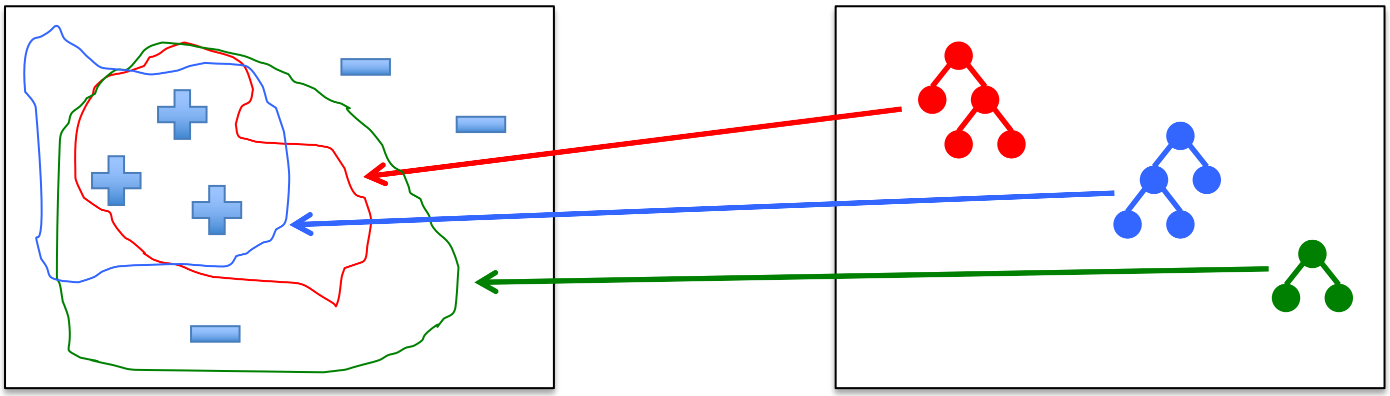
Function Approximation: The Big Picture

Instance Space $\mathcal{X} = \{0,1\}^d$

$\boldsymbol{x} = \langle x_1, x_2, \ldots, x_d \rangle \in \mathcal{X}$

Hypothesis Space

$H = \{h \mid h : \mathcal{X} \mapsto \{0,1\}\}$



if $d = 20$, $|\mathcal{X}| = 2^{20}$

$|h| = 2^{|\mathcal{X}|} = 2^{2^{20}}$

- How many labeled instances are needed to determine which of the $2^{2^{20}}$ hypotheses are correct?
  - All 2^{20} instances in $\mathcal{X}$ must be labeled!
- Generalizing beyond the training data (inductive inference) is impossible unless we add more assumptions (e.g., priors over H)

Bias-Variance Decomposition of Squared Error

- Assume that $y = f(\boldsymbol{x}) + \epsilon$
  - Noise ϵ is sampled from a normal distribution with 0 mean and variance σ^2: $\epsilon \sim N(0, \sigma^2)$
  - Noise lower-bounds the performance (error) we can achieve

- Recall the following objective function:

$$J(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \left(y^{(i)} - h_{\boldsymbol{\theta}} \left(\boldsymbol{x}^{(i)} \right) \right)^2$$

- We can view this as an approximation of the expected value of the squared error: $\mathrm{E} \left(y - h_{\boldsymbol{\theta}} \left(\boldsymbol{x} \right) \right)^2$

Bias-Variance Decomposition of Squared Error

$$
\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{E}[(y - f(\boldsymbol{x}) + f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\qquad + 2\,\mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))(y - f(\boldsymbol{x}))] \\
&= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\qquad + 2\left(\cancel{\mathrm{E}[f(\boldsymbol{x})h_{\boldsymbol{\theta}}(\boldsymbol{x})]} + \cancel{\mathrm{E}[yf(\boldsymbol{x})]} - \cancel{\mathrm{E}[yh_{\boldsymbol{\theta}}(\boldsymbol{x})]} - \cancel{\mathrm{E}[f(\boldsymbol{x})^2]}\right)
\end{aligned}
$$

cancels

cancels

Therefore,

$$
\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{E}[(y - f(\boldsymbol{x}))^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{E}[\epsilon^2] + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]
\end{aligned}
$$

This is actually $\mathrm{var}(\epsilon)$, since mean is 0

Aside:
Definition of Variance

$$\mathrm{var}(z) = \mathrm{E}[(z - \mathrm{E}[z])^2]$$

Bias-Variance Decomposition of Squared Error

$$\begin{aligned}
\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] &= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] + \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2] + \mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\qquad + 2\mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])] \\
&= \mathrm{var}(\epsilon) + \mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2] + \mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] \\
&\qquad + 2\left(\cancel{\mathrm{E}[f(\boldsymbol{x})\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]]} - \cancel{\mathrm{E}[\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]^2]} - \cancel{\mathrm{E}[f(\boldsymbol{x})h_{\boldsymbol{\theta}}(\boldsymbol{x})]} + \cancel{\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})]]}\right)
\end{aligned}$$

cancels cancels

Therefore,

$$\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] = \underbrace{\mathrm{var}(\epsilon)}_{\text{noise}} + \underbrace{\mathrm{E}[(f(\boldsymbol{x}) - \mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})])^2]}_{\text{bias}} + \underbrace{\mathrm{E}[(\mathrm{E}[h_{\boldsymbol{\theta}}(\boldsymbol{x})] - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2]}_{\text{variance}}$$

$$\mathrm{E}[(y - h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2] = \mathrm{bias}(h_{\boldsymbol{\theta}}(\boldsymbol{x}))^2 + \mathrm{var}(h_{\boldsymbol{\theta}}(\boldsymbol{x})) + \sigma^2$$

Regularization

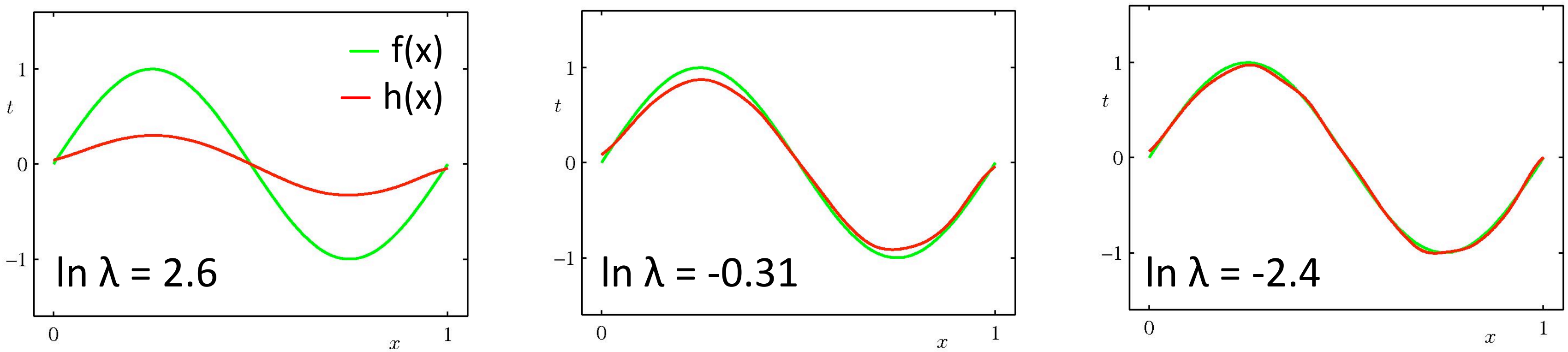
- Linear regression objective function

$$J(\boldsymbol{\theta}) = \underbrace{\frac{1}{2n}\sum_{i=1}^{n}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}^{(i)}\right) - y^{(i)}\right)^2}_{\text{model fit to data}} + \underbrace{\frac{\lambda}{2}\sum_{j=1}^{d}\theta_j^2}_{\text{regularization}}$$

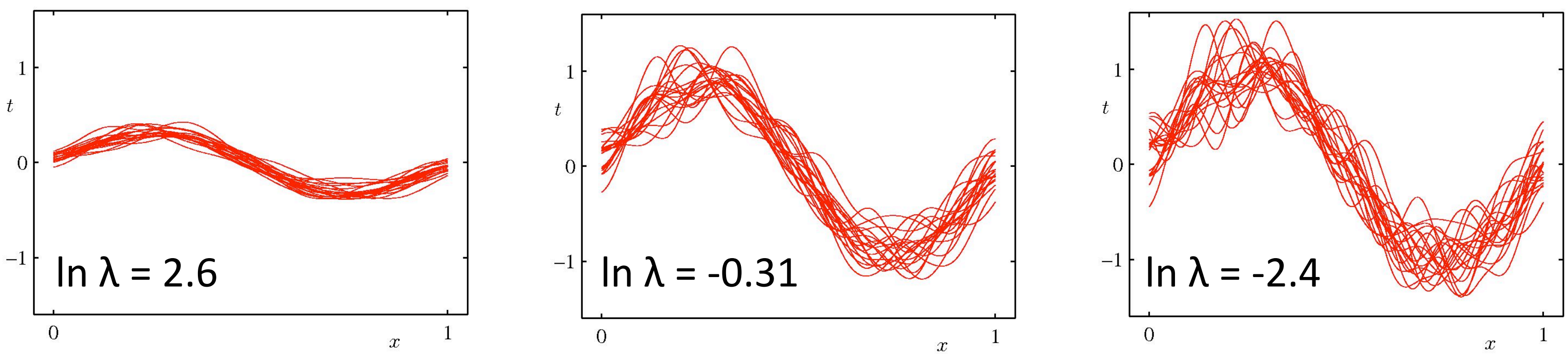
  – λ is the regularization parameter ($\lambda \geq 0$)

Illustration of Bias-Variance

high ← Bias → low

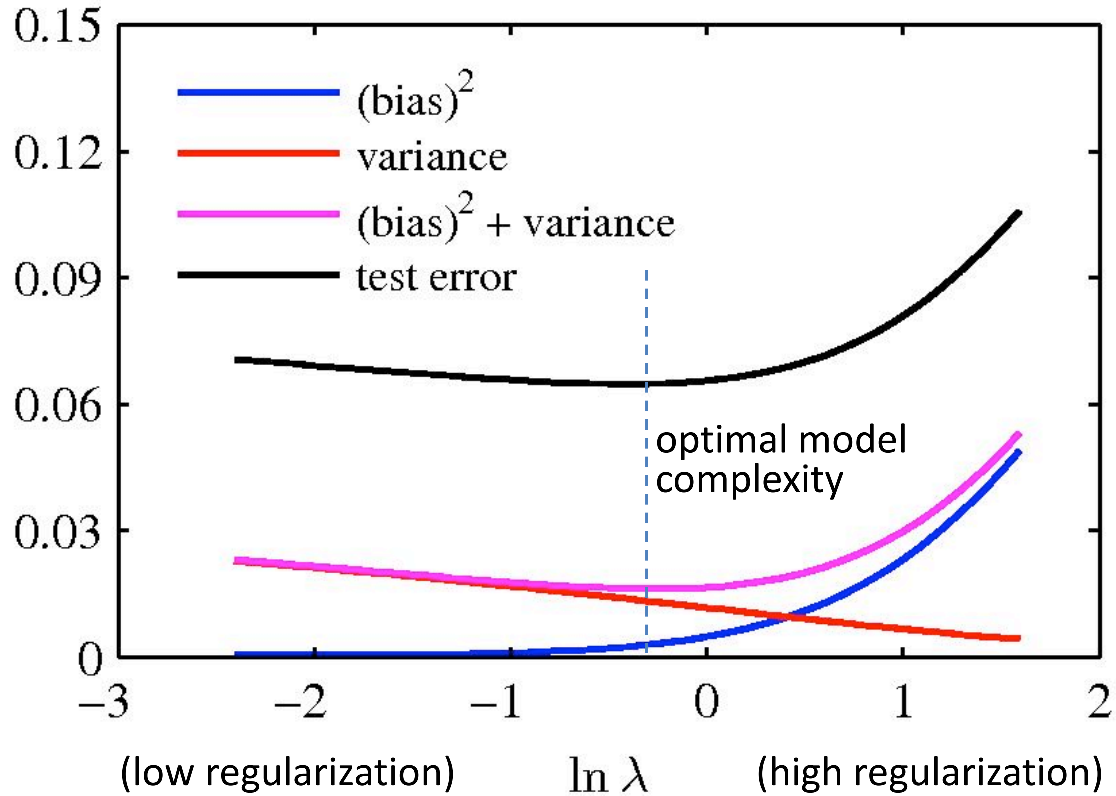


high regularization ↔ low regularization



low ← Variance → high

Illustration of Bias-Variance



- Reducing training error drives down bias, but ignores variance

Figure provided by by Max Welling