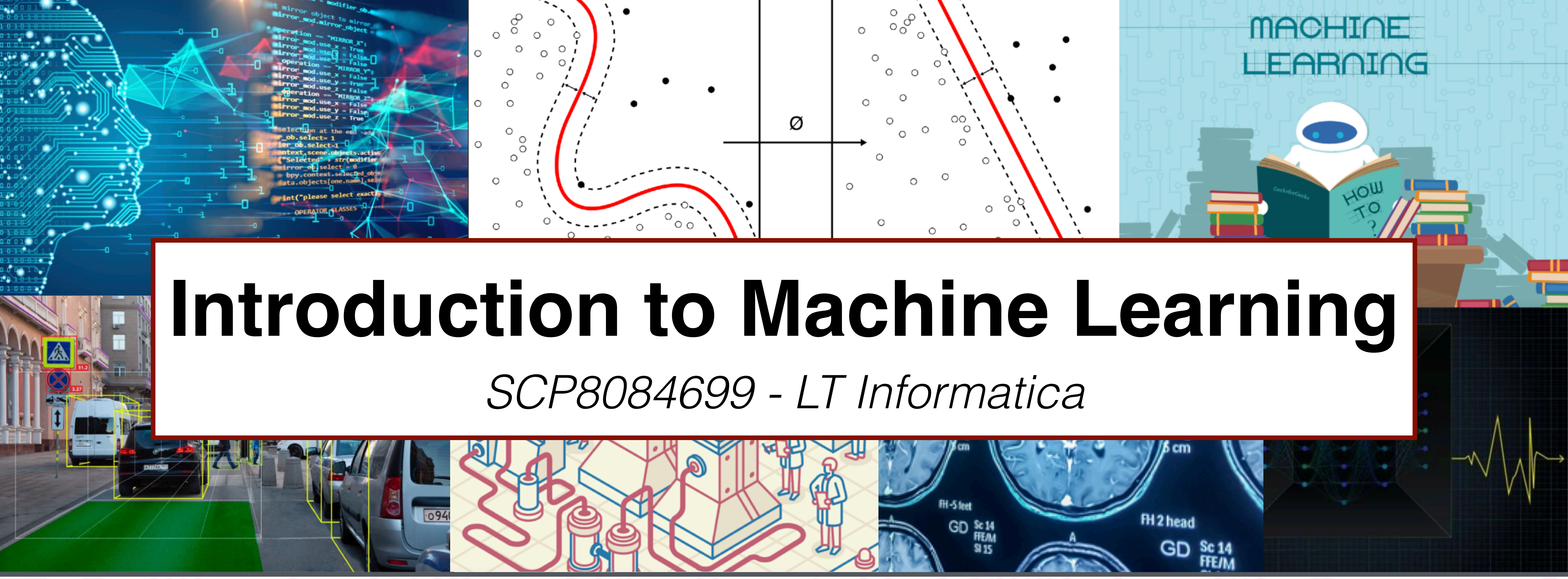


Introduction to Machine Learning

SCP8084699 - LT Informatica

ML System Design, Diagnoses and Learning Curves

Prof. Lamberto Ballan

DIPARTIMENTO MATEMATICA
Department of Mathematics "Tullio Levi-Civita"

What we will learn today?

- A bit more on model selection and evaluation: advice for applying machine learning
- Diagnosing machine learning models

Recap: Supervised Learning

- Classification (discrete) *vs* Regression (real-valued output)

$\{(x^{(i)}, y^{(i)})\}$

Size in feet² (x)	Price ($) in 1K's (y)
2104	460
1416	232
1534	315
852	178
…	…

Traini... t

...ning Algorithm

$x \longrightarrow \boxed{h} \longrightarrow y$

Hypothesis Space H

Looking for a good
Bias-Variance Tradeoff

h approximates the unknown target f

$$h \sim f: X \rightarrow Y$$

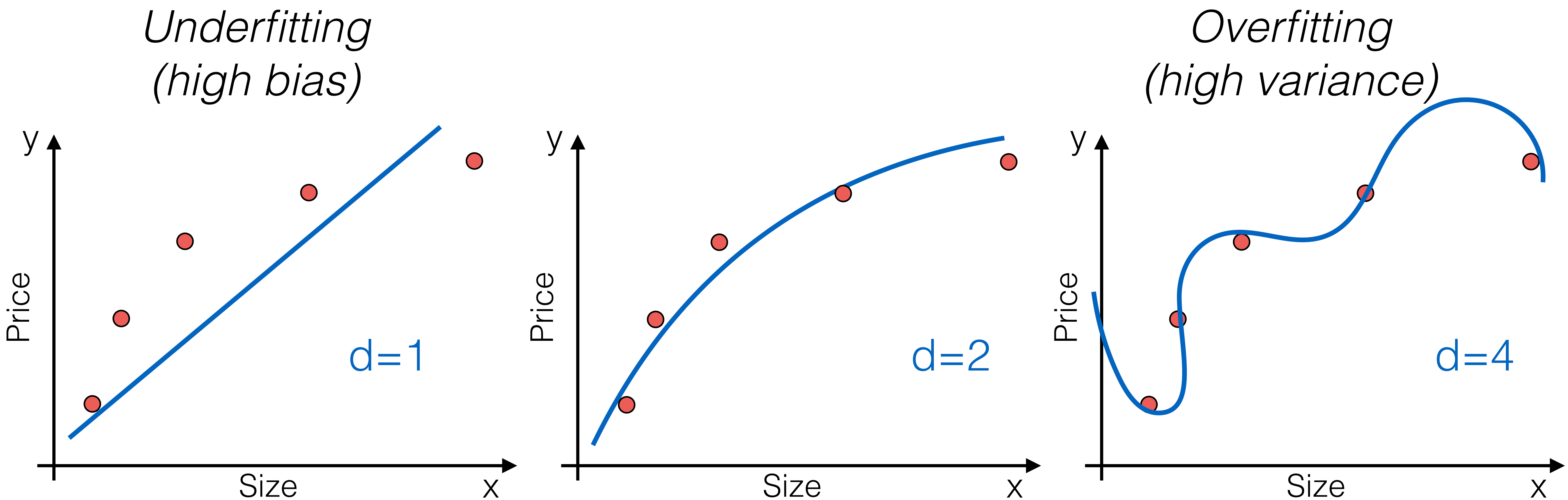
Model Selection and Evaluation

- **Hold-out**: we keep a subset of v samples from the training set (the validation set) to evaluate our model
  - A classifier/regressor is trained on m-v samples
  - Parameters are optimized on the *training*-*validation* sets: then you should evaluate performances on the *test* set
  - Size (cardinality) of training+validation sets should be greater than test set, e.g. 70%,15%,15%
- ***k*-fold cross validation**: iterate on *k* disjoint subsets
- Given a task, pick the "right" evaluation metric

Diagnosing bias *vs* variance

- If your learning model doesn't work as expected, almost all the time it will be because you have either a *high bias* problem or a *high variance* problem
  - How to figure out what's happening (in practice)?
  - What can we do to fix/alleviate the problem?

Diagnosing bias vs variance



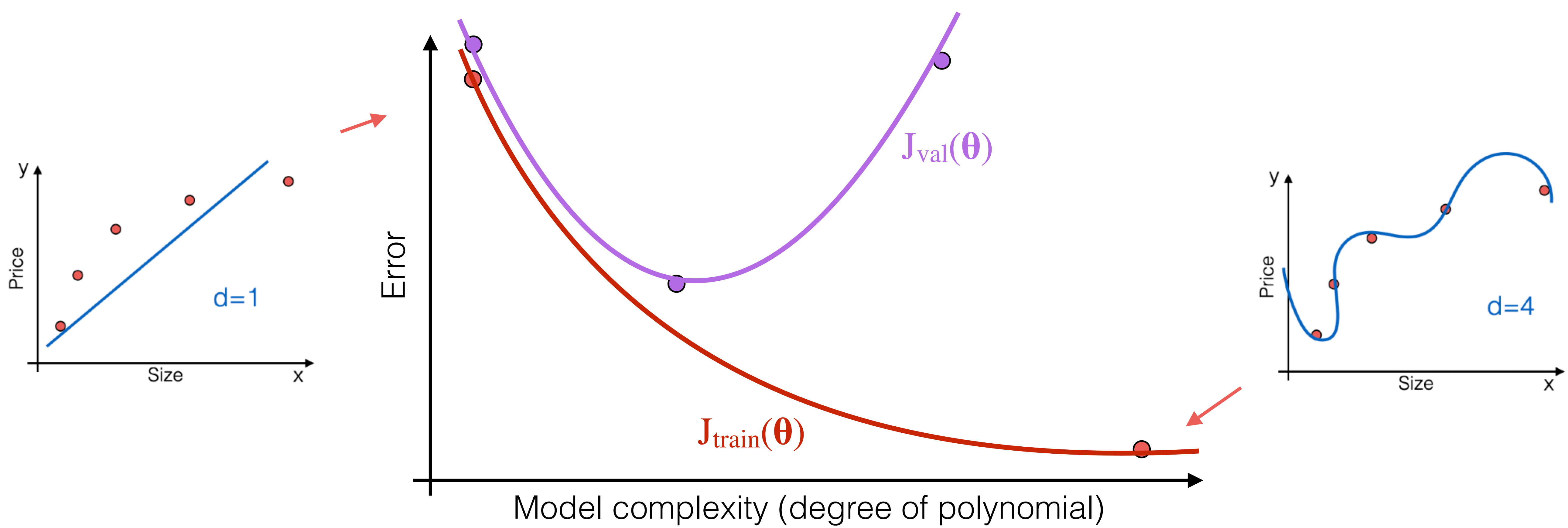
$h_\theta(x) = \theta_0 + \theta_1 x$ $\quad$ $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2$ $\quad$ $\theta_0+\theta_1 x+\theta_2 x^2+\theta_3 x^3+\theta_4 x^4$

- We can now look again at this example taking into account hold-out and bias-variance tradeoff

Adapted from slide by A. Ng

Diagnosing bias vs variance

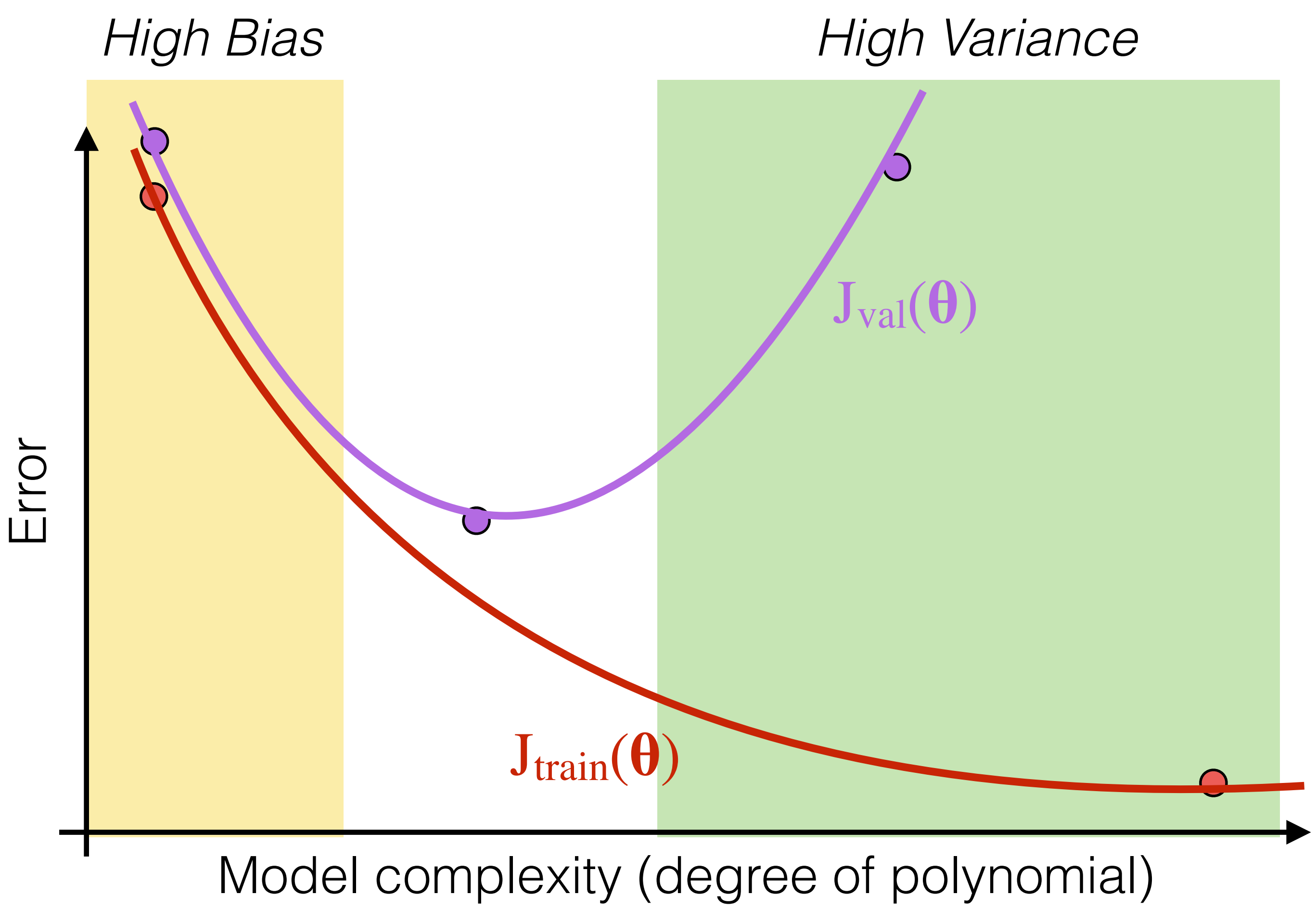
- "Measuring" bias vs variance:
  - Training Error: $J_{train}(\boldsymbol{\theta}) = \dfrac{1}{2m_t} \sum_{i=1}^{m_t} (h_\theta(x^{(i)}) - y^{(i)})^2$
  - Validation Error: $J_{val}(\boldsymbol{\theta}) = \dfrac{1}{2m_v} \sum_{i=1}^{m_v} (h_\theta(x^{(i)}) - y^{(i)})^2$



Adapted from slide by A. Ng

Diagnosing bias vs variance

- Our learning model doesn't work as expected; is it a bias problem or a variance problem?



High bias (underfit):

$J_{train}(\theta)$ will be high

$J_{val}(\theta) \approx J_{train}(\theta)$

High variance (overfit):

$J_{train}(\theta)$ will be low

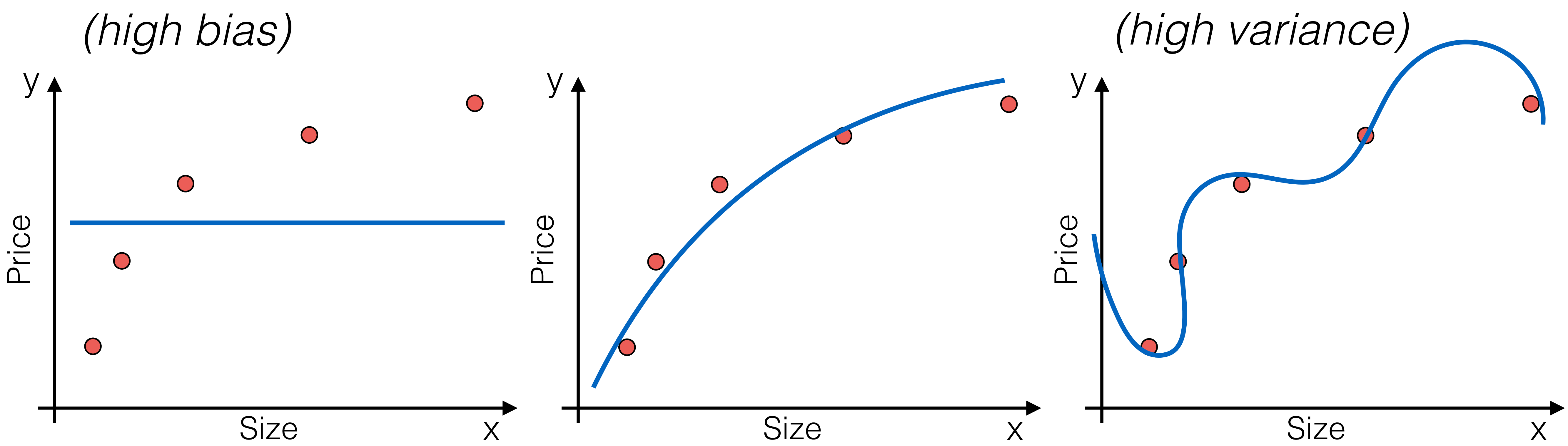
$J_{val}(\theta) \gg J_{train}(\theta)$

Diagnosing bias vs variance

- What's the contribution of regularization?

$$\min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$



Large λ $h_\theta(x) \approx \theta_0$

Intermediate λ

Small λ (i.e. $\lambda \approx 0$)

Diagnosing bias vs variance

- Choosing the regularization parameter λ:

$$\min_{\boldsymbol{\theta}} \frac{1}{2m}\left[\sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2 \right]$$

$$h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

- Note: our definition of J_{train}, J_{val}, J_{test} don't change
  - Training Error: $J_{train}(\boldsymbol{\theta}) = \frac{1}{2m_t} \sum_{i=1}^{m_t} (h_\theta(x^{(i)}) - y^{(i)})^2$
  - Validation Error: $J_{val}(\boldsymbol{\theta}) = \frac{1}{2m_v} \sum_{i=1}^{m_v} (h_\theta(x^{(i)}) - y^{(i)})^2$
  - Test Error: $J_{test}(\boldsymbol{\theta}) = \frac{1}{2m_e} \sum_{i=1}^{m_e} (h_\theta(x^{(i)}) - y^{(i)})^2$

Diagnosing bias vs variance

- Choosing the regularization parameter λ:

$$\min_{\boldsymbol{\theta}} \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda\sum_{j=1}^{n}\theta_j^2\right]$$

Model: $h_\theta(x) = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$

Model Selection

1: try $\lambda=0$ $\longrightarrow$ $\min_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \boldsymbol{\theta}^{(1)}$ $\longrightarrow$ $J_{val}(\boldsymbol{\theta}^{(1)})$

2: try $\lambda=0.01$ $\longrightarrow$ $\boldsymbol{\theta}^{(2)}$ $\longrightarrow$ $J_{val}(\boldsymbol{\theta}^{(2)})$

3: try $\lambda=0.02$ $\longrightarrow$ $\boldsymbol{\theta}^{(3)}$ $\longrightarrow$ $\boxed{J_{val}(\boldsymbol{\theta}^{(3)})}$ *(lowest)*

4: try $\lambda=0.04$ $\longrightarrow$ $\boldsymbol{\theta}^{(4)}$ $\longrightarrow$ $J_{val}(\boldsymbol{\theta}^{(4)})$

5: try $\lambda=0.08$ $\longrightarrow$ $\boldsymbol{\theta}^{(5)}$ $\longrightarrow$ $J_{val}(\boldsymbol{\theta}^{(5)})$

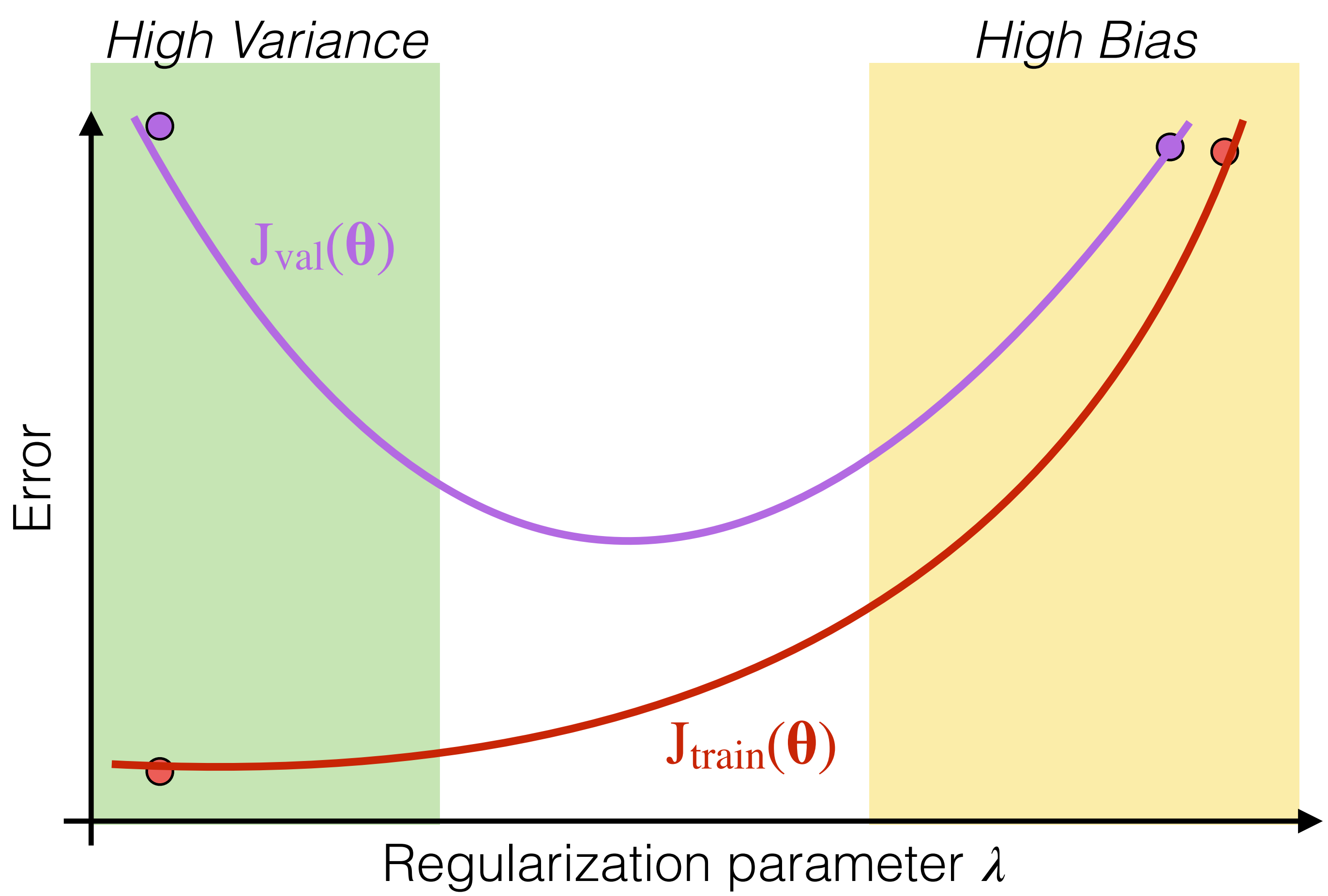
⋮

12: try $\lambda\approx10$ $\longrightarrow$ $\boldsymbol{\theta}^{(12)}$ $\longrightarrow$ $J_{val}(\boldsymbol{\theta}^{(12)})$

Diagnosing bias vs variance

- Bias/Variance as a function of the parameter λ:

$$\min_{\theta} \frac{1}{2m}\left[\sum_{i=1}^{m}(h_\theta(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^{n} \theta_j^2\right]$$



$$J_{train}(\theta) = \frac{1}{2m_t}\sum_{i=1}^{m_t}(h_\theta(x^{(i)}) - y^{(i)})^2$$

$$J_{val}(\theta) = \frac{1}{2m_v}\sum_{i=1}^{m_v}(h_\theta(x^{(i)}) - y^{(i)})^2$$

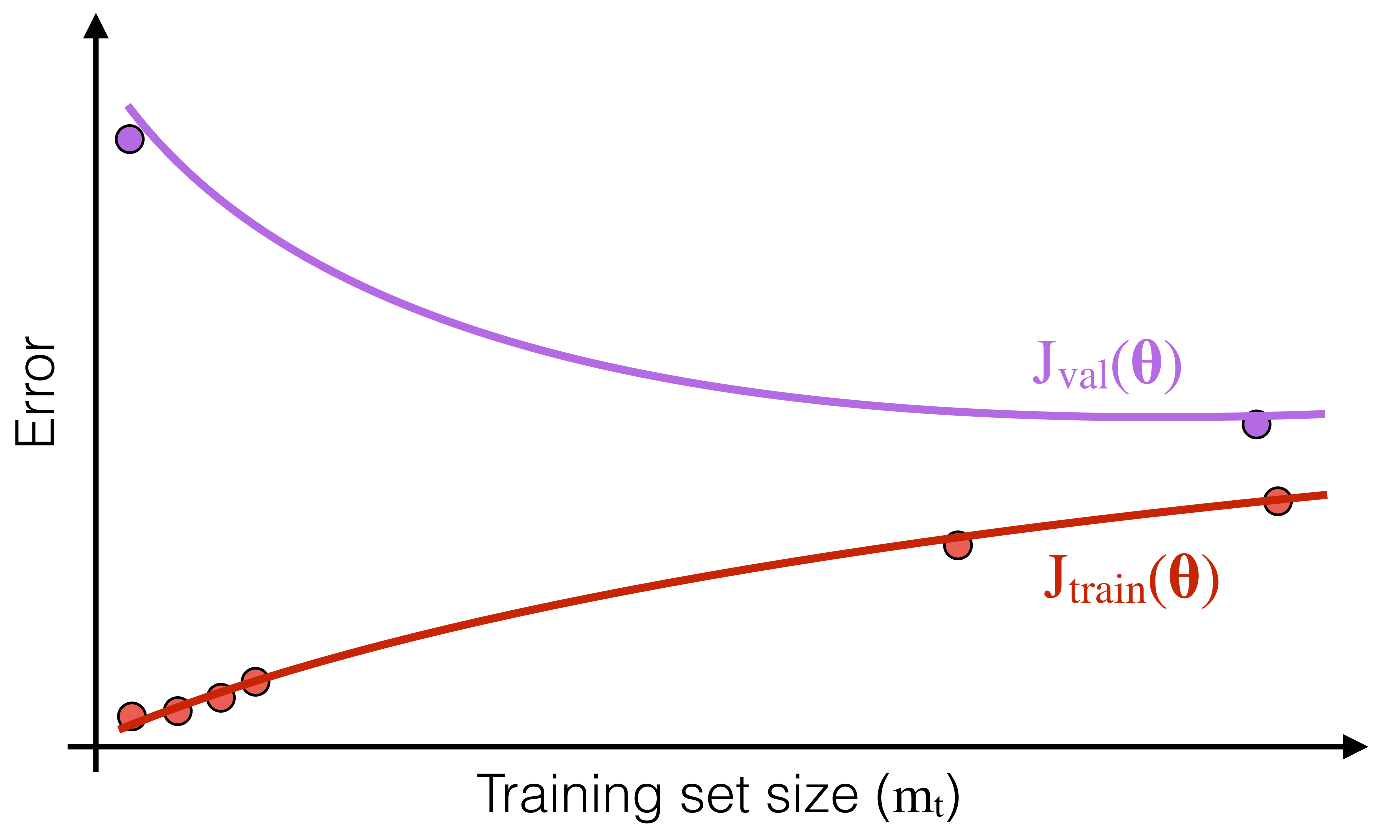
Diagnosing bias vs variance

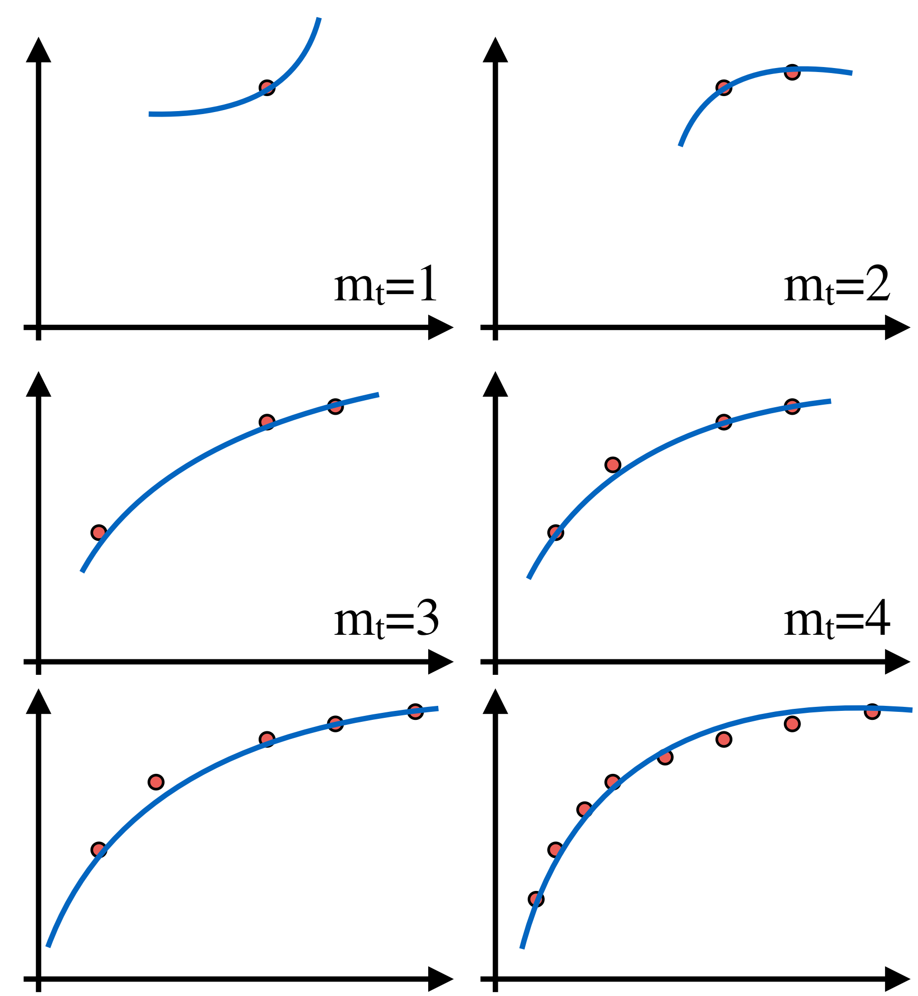
- By now you have seen bias and variance from a lot of different perspectives
- Let’s now take all the insights we have gone through in order to build a “diagnostic tool” for ML systems

Learning Curves

- Learning curves can be used to diagnose if a model may be suffering from bias, variance or a bit of both

$$J_{train}(\boldsymbol{\theta}) = \frac{1}{2m_t} \sum_{i=1}^{m_t} (h_\theta(x^{(i)}) - y^{(i)})^2 \qquad J_{val}(\boldsymbol{\theta}) = \frac{1}{2m_v} \sum_{i=1}^{m_v} (h_\theta(x^{(i)}) - y^{(i)})^2$$

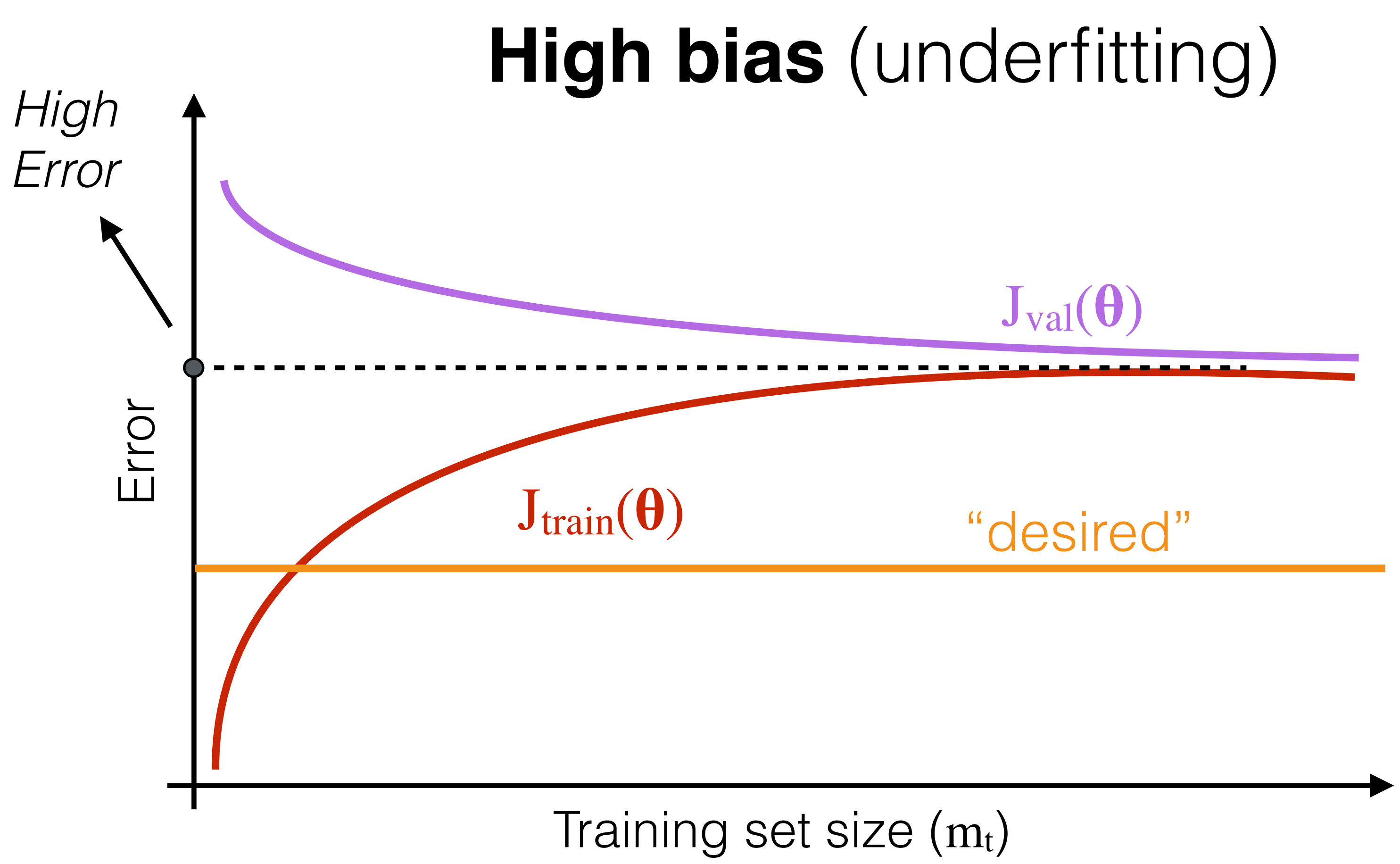


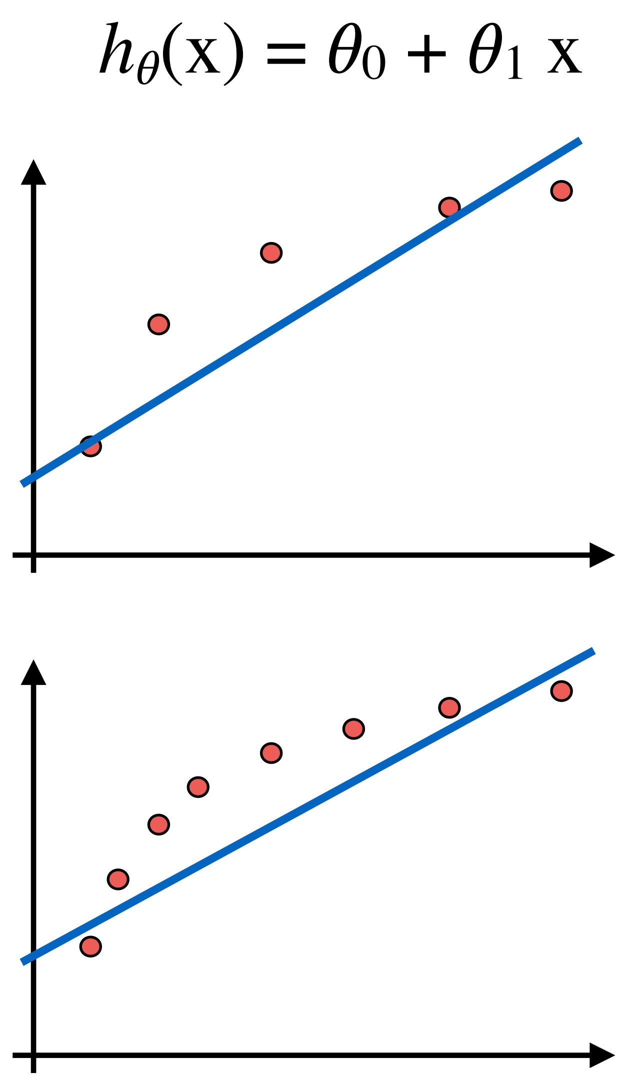


$h_\theta(x) = \theta_0 + \theta_1\, x + \theta_2\, x^2$

Learning Curves

- That's the general intuition… but what's about bias and variance problems?



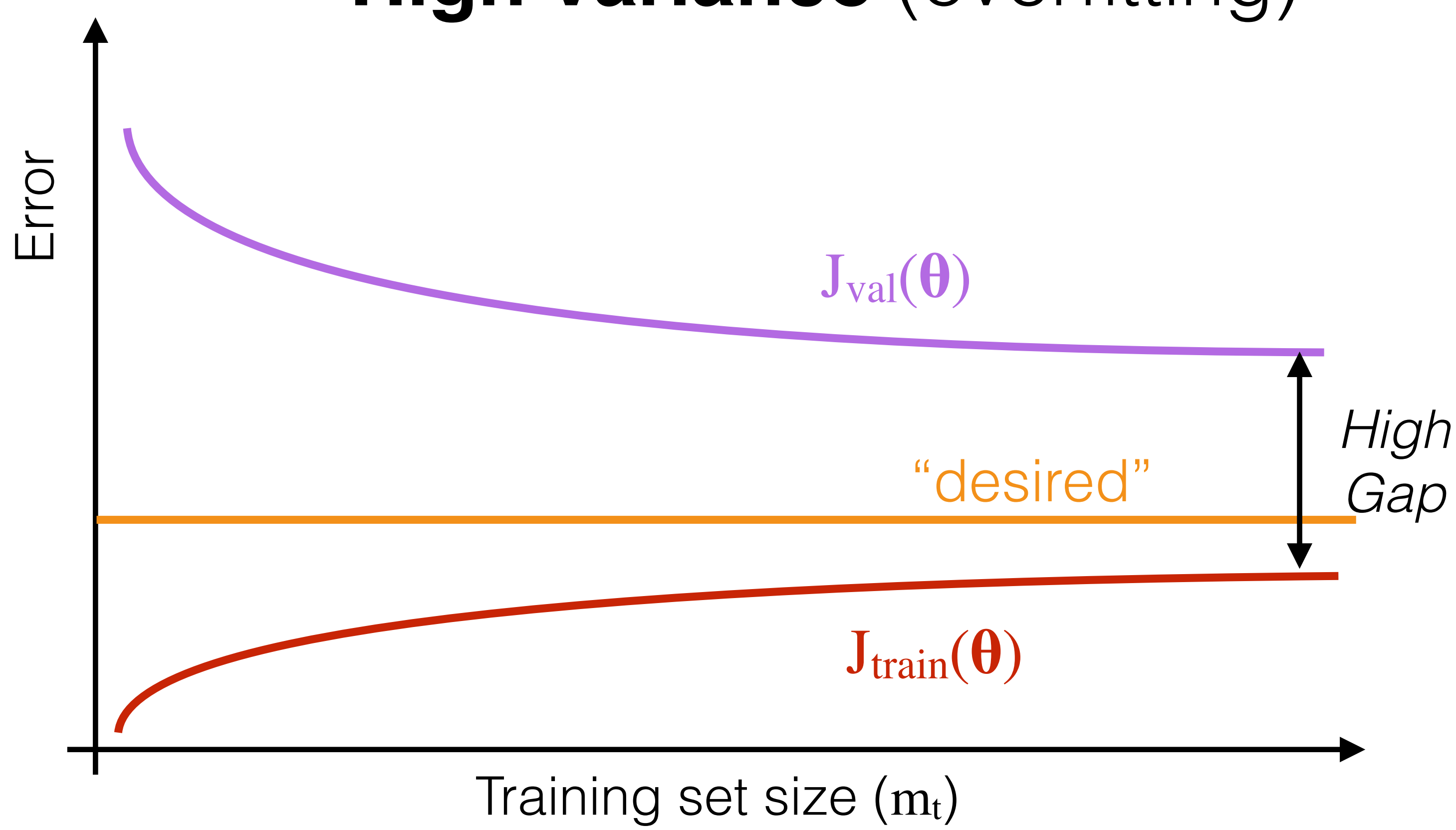


Note: in case of high bias, getting more training data will not help much

Learning Curves

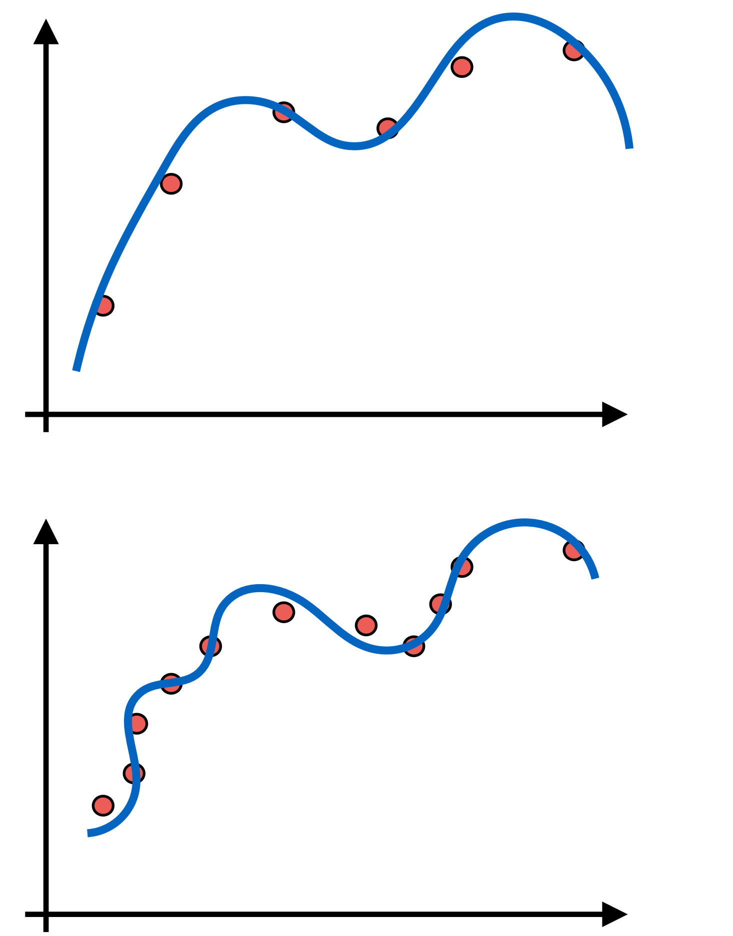
- That's the general intuition… but what's about bias and variance problems?

High variance (overfitting)



$h_\theta(x) = \theta_0 + \theta_1 x + \ldots + \theta_{50} x^{50}$

(and small λ)

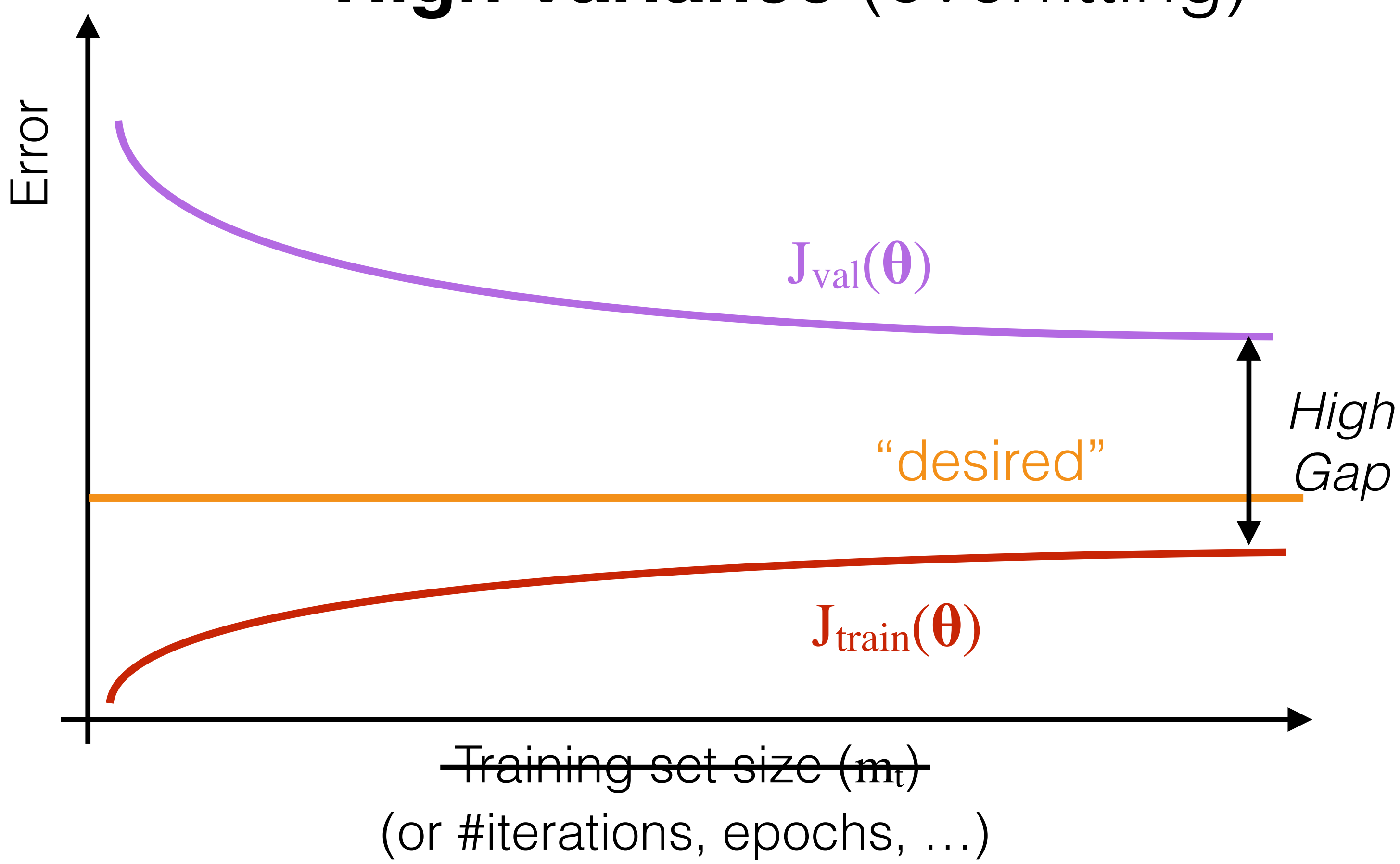


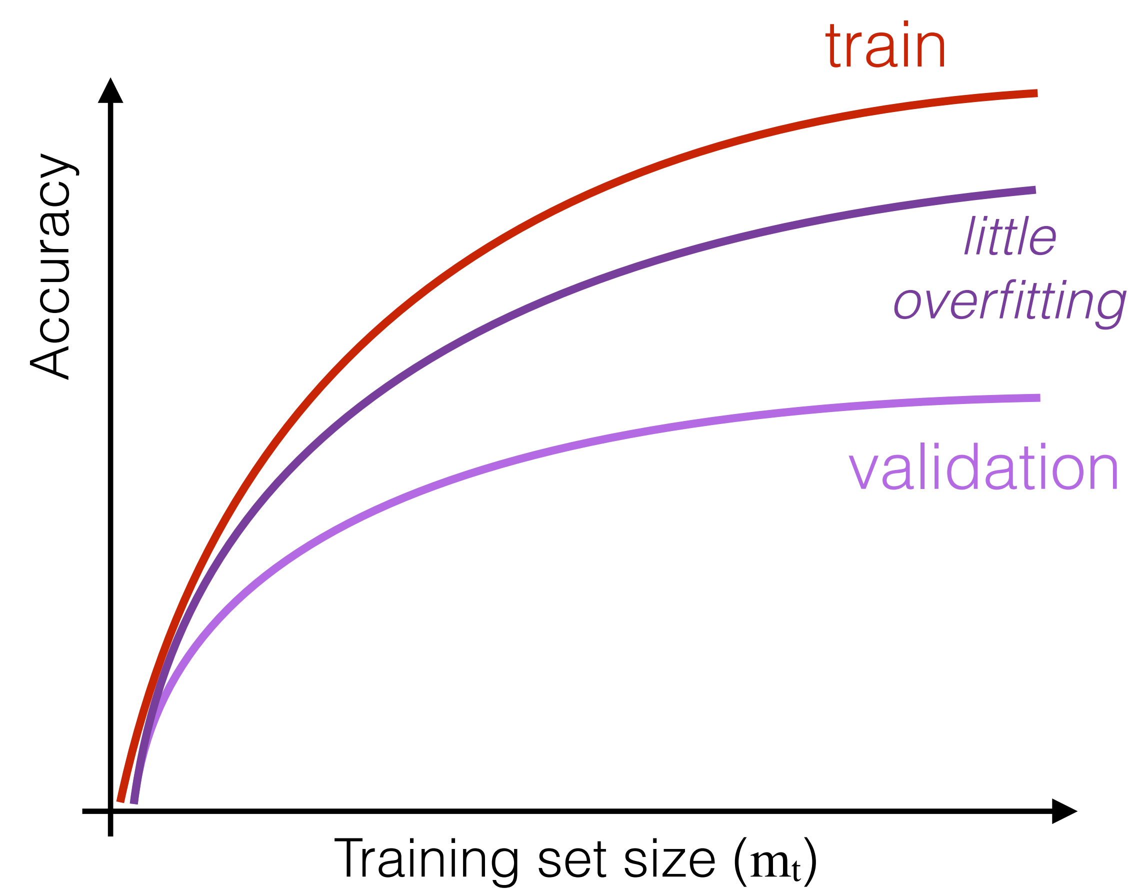
Note: in case of high variance, getting more training data is likely to help

Learning Curves

- You can compute learning curves w.r.t. different "dimensions" (e.g. evaluation measures, no. samples)

High variance (overfitting)





What to do next

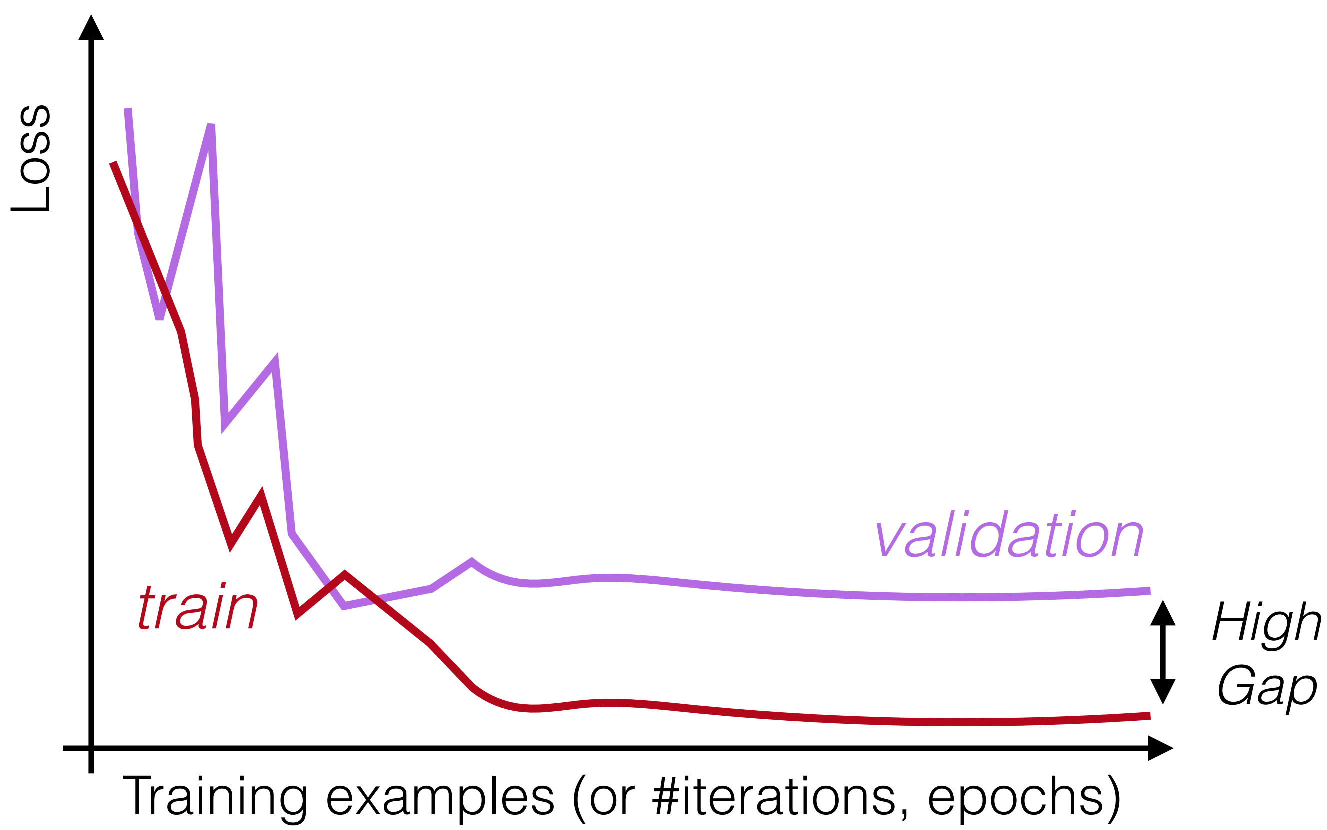
- Debugging (and babysitting) a learning algorithm:
  - Suppose you have implemented a regularized linear regression model for predicting housing prices
  - It doesn't work on new data; what should you do next?
    - You can get more training data → *Fixes high variance*
    - Try smaller set of features → *Fixes high variance*
    - Try getting more features → *Fixes high bias*
    - Try adding complexity to the model (e.g. polynomial features) → *Fixes high bias*
    - Try decreasing λ → *Fixes high bias*
    - Try increasing λ → *Fixes high variance*

Adapted from slide by A. Ng

Diagnosing our datasets

- Learning curves can be also used to diagnose the quality of our training/validation sets

Unrepresentative Training Set

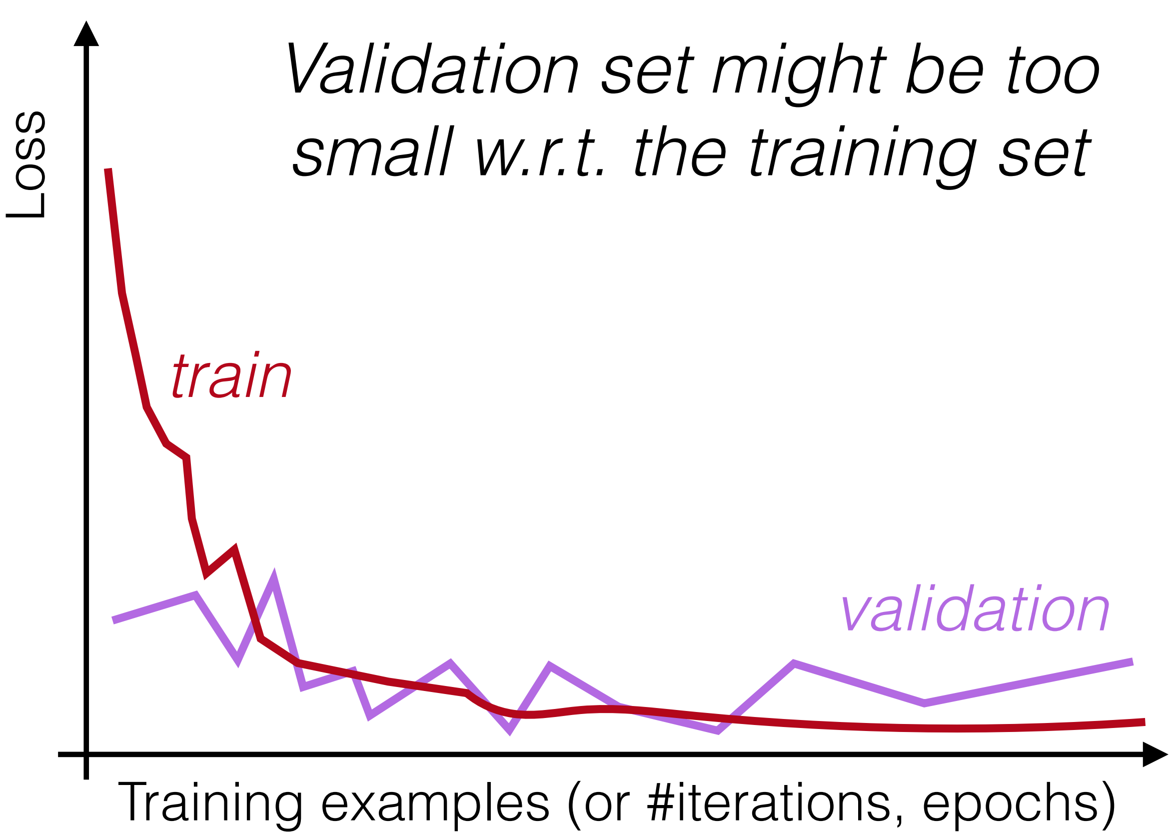


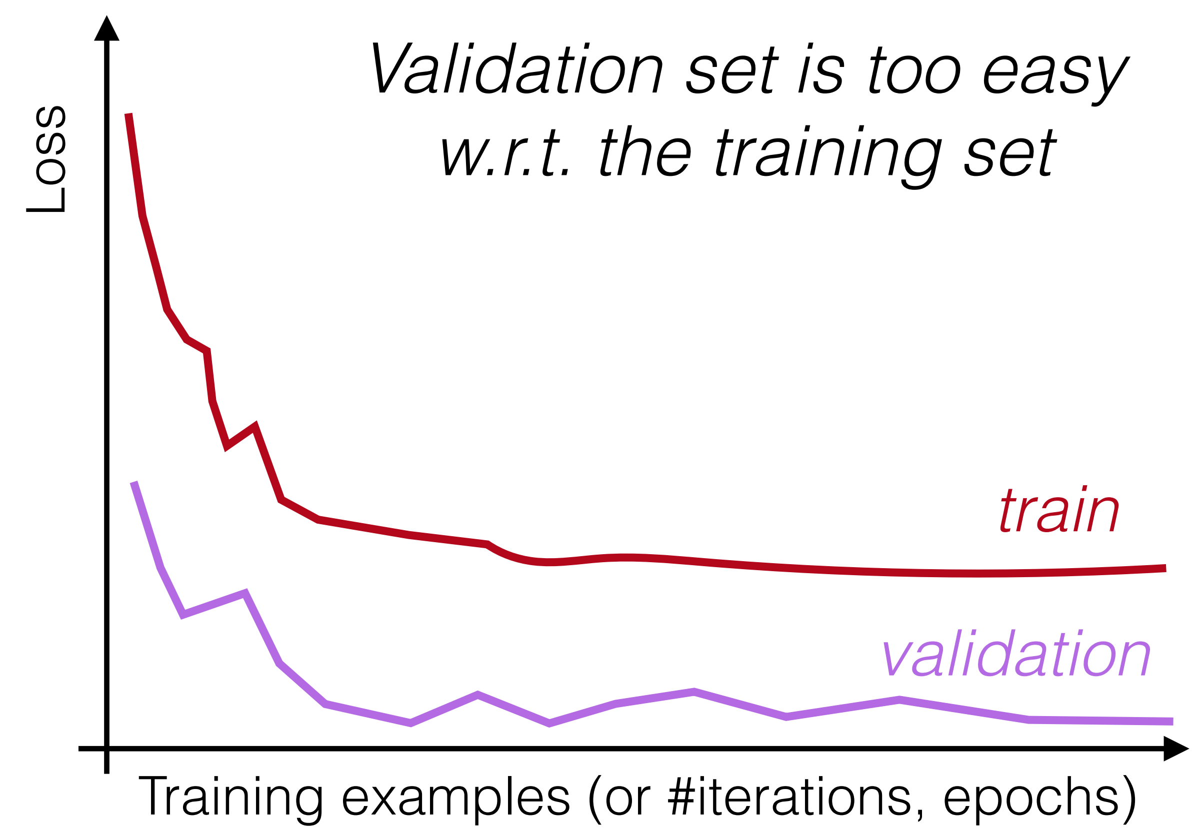
- The training set does not provide sufficient information to learn the problem
- It may occur if the training set has too few examples as compared to the validation set

Diagnosing our datasets

- Learning curves can be also used to diagnose the quality of our training/validation sets

Unrepresentative Validation Set





Contact

- **Office:** Torre Archimede 3CD, room 320
- **Office hours** (ricevimento)**:** Monday 11:00-13:00

✉ lamberto.ballan@unipd.it

🏠 http://www.lambertoballan.net

🏠 http://vimp.math.unipd.it

@ twitter.com/lambertoballan