UNIVERSITY OF AMSTERDAM
Informatics Institute

AMLAB Amsterdam Machine Learning Lab

Machine Learning 1

Lecture 13.2 - Combining Models
Bootstrapping and Feature Bagging - Recap
Bias-Variance Decomposition

Erik Bekkers

(Bishop 14.2)

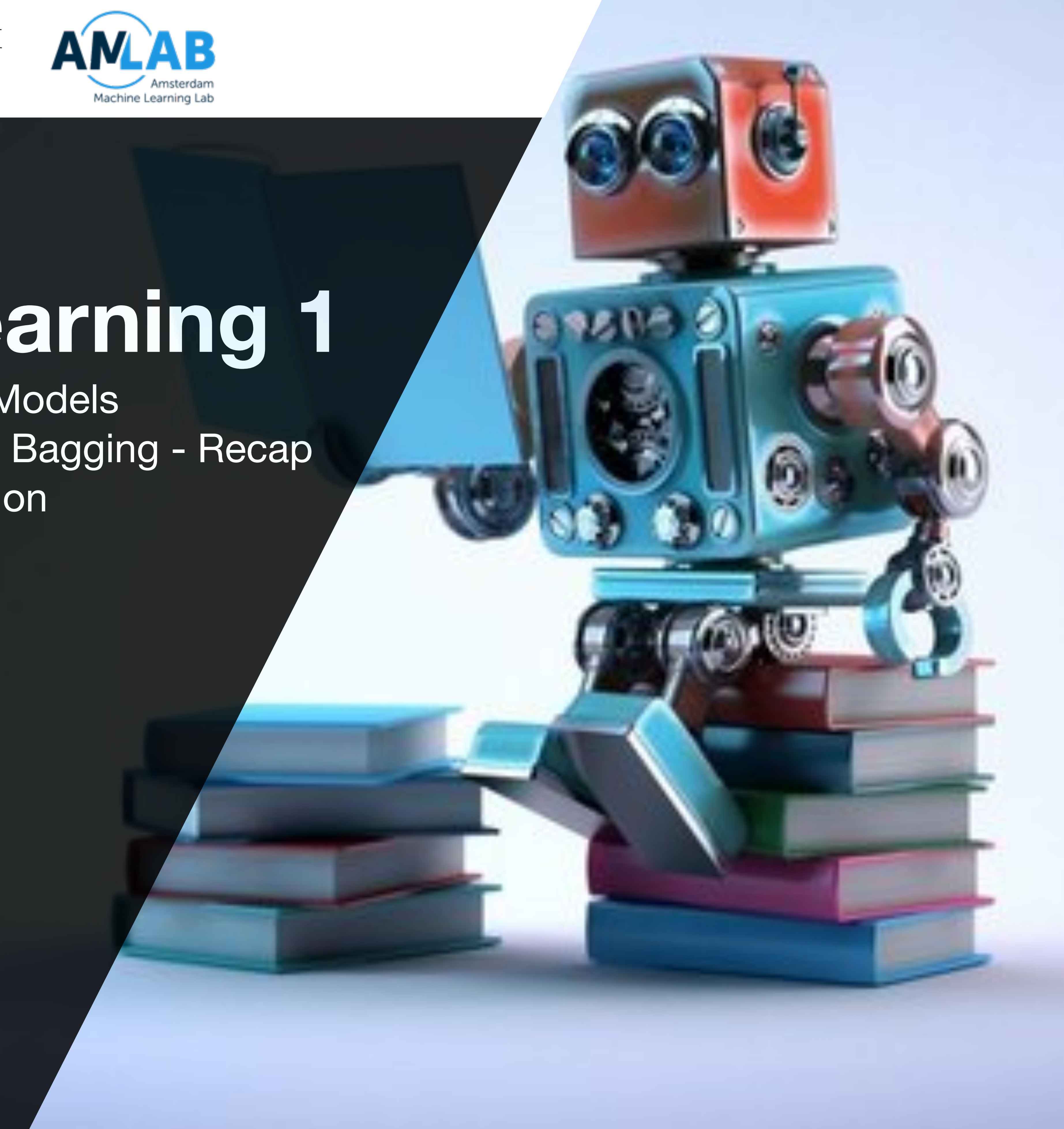

Image credit: Kirillm | Getty Images

Slide credits: Patrick Forré and Rianne van den Berg

Regression with GP’s

- Combining models: (Bishop 4.1-4.4)
  - Bayesian model averaging vs. model combination methods
  - **Committees**:
    - **Bootstrap aggregation**
    - **Random subspace methods**
    - Boosting
  - Decision trees
  - Random forests

Constructing committees

- Simplest way to construct a committee is by averaging predictions of a set of individual models

- Remember the bias variance trade-off: model error decomposes into two components

  - Bias: arises from the difference between model and the ground truth function that needs to predicted

  - Variance: represents the sensitivity of a model to the individual datapoints that it was trained on

Bias-Variance Decomposition: Example

- Generate L ($=100$) datasets of N ($=25$) points:

$$x \sim U(0,1)$$

$$t = \sin(2\pi x) + \epsilon \qquad \epsilon \in N(0, \alpha^{-1})$$

$$\mathbb{E}[t \,|\, x] = \sin(2\pi x)$$

- L predictions with 24 Gaussian basis functions:

$$E_D = \frac{1}{2}\sum_{i=1}^{N}\{t_n - \mathbf{w}^T\phi(x)\}^2 + \frac{\lambda}{2}\mathbf{w}^T\mathbf{w}$$

$$y^{(l)}(x) = (\mathbf{w}^{(l)})^T\phi(x)$$

$$\mathbb{E}_D[y_D(x)] = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

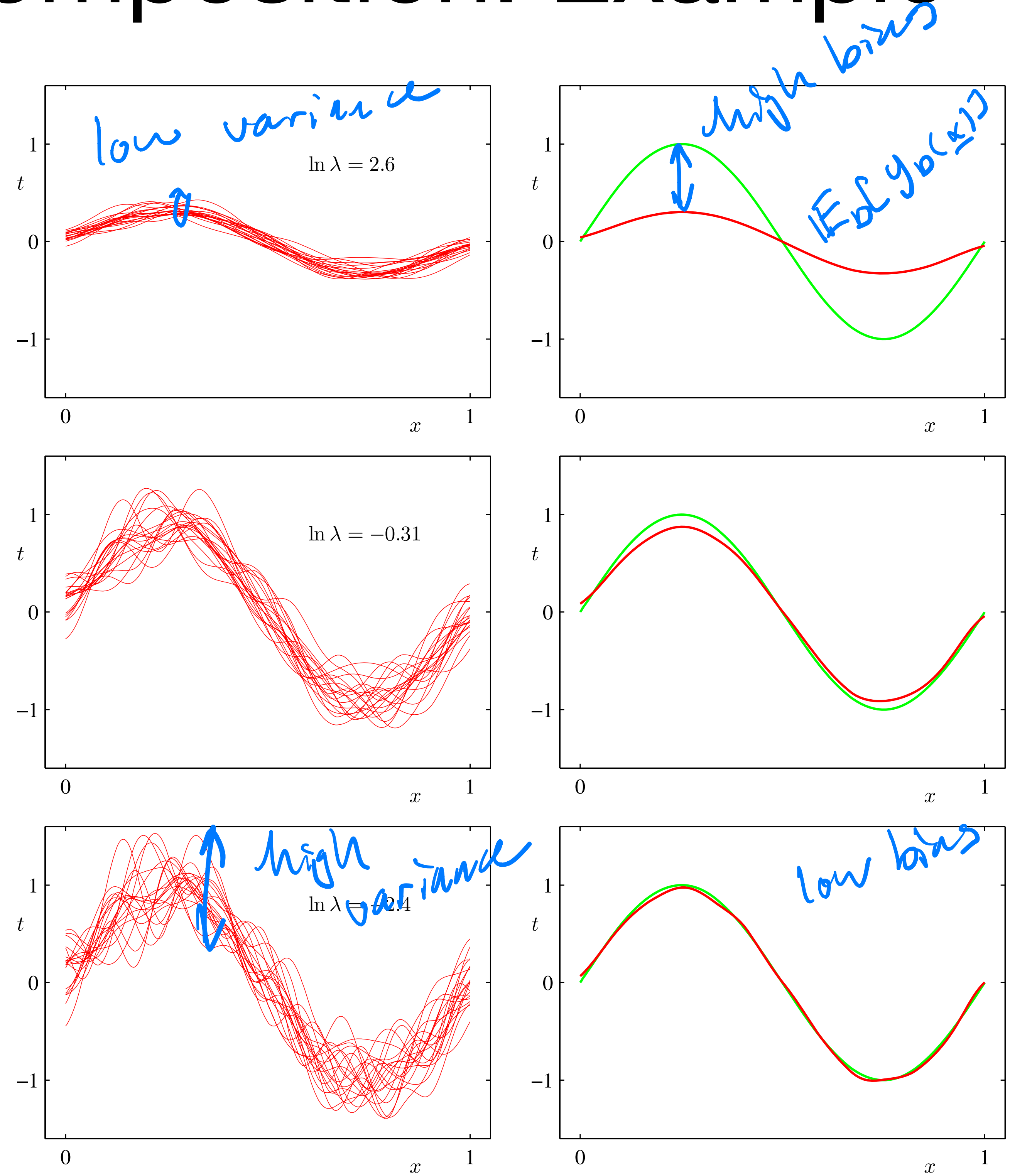

Figure: bias-variance decomposition (Bishop 3.5)

Averaging predictions from different models

- When we average models trained on different datasets, the contribution of the variance reduces

- When we average a set of low-bias models (corresponding to complex models such as high-order polynomials), we obtained accurate predictions!

- However, in practice we only have one single dataset!

- One way to introduce variability between different models within the committee: bootstrap datasets.

Committees: bootstrapping datasets

- Suppose your original dataset consists of N data points:

$$\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right]^T$$

each X_b is a new dataset of size N

- Create B new datasets $\{\mathbf{X}_1, \ldots, \mathbf{X}_B\}$ by drawing N points at random from $\mathbf{X}$, with replacement.

- Some data points will occur multiple times in $\mathbf{X}_b$

- Some data points will be absent from $\mathbf{X}_b$

Regression with B bootstrap datasets

- We have generated B bootstrap datasets $\{\mathbf{X}_1, \ldots, \mathbf{X}_B\}$
- Use each $\mathbf{X}_b$ to train a separate model $y_b(\mathbf{x})$
- The committees prediction $y_{\mathrm{COM}} = \frac{1}{B}\sum_{b=1}^{B} y_b(\mathbf{x})$
- **This is called bootstrap aggregation/bagging!**
- Suppose the ground truth function that we need to predict is $h(\mathbf{x})$
- The prediction of each individual model: $y_b(\mathbf{x}) = h(\mathbf{x}) + \epsilon_b(\mathbf{x})$
- Error of model b: $\epsilon_b(\mathbf{x})$

Bootstrap aggregation

$$\mathbb{E}\left[\frac{1}{B^2}\left(\sum_{b=1}^{B} \epsilon_b(x)\right)\left(\sum_{b'=1}^{B} \epsilon_{b'}(x)\right)\right]$$

$$= \frac{1}{B^2}\sum_{b=1}^{B}\sum_{b'=1}^{B} \mathbb{E}\left[\delta_{b=b'}\, \epsilon_b(x)\, \epsilon_{b'}(x)\right]$$

$$= \frac{1}{B^2}\sum_{b=1}^{B} \mathbb{E}\left[\epsilon_b(x)^2\right]$$

$$= \frac{1}{B} E_{AV}$$

- The average sum-of-squares error for model b:

$$\mathbb{E}_{\mathbf{x}}[\{y_b(\mathbf{x}) - h(\mathbf{x})\}^2] = \mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})^2]$$

- The average error made by the b models individually:

$$E_{\mathrm{AV}} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})^2]$$

- The expected error of the committee $y_{\mathrm{COM}} = \frac{1}{B}\sum_{b=1}^{B} y_b(\mathbf{x})$

$\left(h(x) = \frac{1}{B}\sum_b h(x)\right)$

$$E_{\mathrm{COM}} = \mathbb{E}_{\mathbf{x}}\left[\left\{\underbrace{\frac{1}{B}\sum_{b=1}^{B} y_b(\mathbf{x})}_{y_{\mathrm{COM}}} - h(\mathbf{x})\right\}^2\right] = \mathbb{E}_{\mathbf{x}}\left[\left\{\frac{1}{B}\sum_{b=1}^{B} \epsilon_b(\mathbf{x})\right\}^2\right]$$

- If we assume $\mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})] = 0$ and $\mathrm{cov}[\epsilon_b(\mathbf{x}), \epsilon_{b'}(\mathbf{x})] = 0$ for $b' \neq b$ then

$$\mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})\epsilon_{b'}(\mathbf{x})] = 0 \qquad E_{\mathrm{COM}} = \frac{1}{B}E_{\mathrm{AV}}$$

Bootstrap aggregation

- If we assume $\mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})] = 0$ and $\mathbb{E}_{\mathbf{x}}[\epsilon_b(\mathbf{x})\epsilon_{b'}(\mathbf{x})] = 0$ for $b' \neq b$

$$E_{\mathrm{COM}} = \frac{1}{B} E_{\mathrm{AV}}$$

- Seems like the average error of a model due to the variance can be reduced by a factor B if we average B versions of the model…
- However, we assumed that error due to individual models are uncorrelated!
- In practice, errors are highly correlated (the bootstrap datasets are not independent)
- But $E_{\mathrm{COM}} \leq E_{\mathrm{AV}}$ even for correlated errors!
- Strategy: choose B models with low-bias (complex models that can overfit), bootstrap aggregated model will have lower error, than the average error of the individual models.

Committees: feature bagging

- Feature bagging: **sample a subset of features** of length $r < D$ for each learner $\underline{x} = (x_1, x_2, x_3, x_4, x_5, x_6)^T \in \mathbb{R}^D$

y_1

y_2

y_3

- Also called '**random subspace method**'
- Works especially well if features are uncorrelated
- Causes learners to not over-focus on features that are overly predictive for training set but do not generalize to new data
- So feature bagging works well if the number of features is much larger than the number of training points
- Decisions trees with bootstrapping and random subspaces -> random forests