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 be predicted.

    - **Variance**: represents the sensitivity of a model to the individual datapoints that it was trained on.
Bias-Variance Decomposition: Example

- Generate $L$ datasets of $N$ points:

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

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

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

- $L$ predictions with 24 Gaussian basis functions:

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

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

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

**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:

\[
X = [x_1, \ldots, x_N]^T
\]

- Create \( B \) new datasets \( \{X_1, \ldots, X_B\} \) by drawing \( N \) points at random from \( X \), with replacement.

- Some data points will occur multiple times in \( X_b \)

- Some data points will be absent from \( X_b \)
Regression with $B$ bootstrap datasets

- We have generated $B$ bootstrap datasets $\{X_1, \ldots, X_B\}$

- Use each $X_b$ to train a separate model $y_b(x)$

- The committees prediction $y_{\text{COM}} = \frac{1}{B} \sum_{b=1}^{B} y_b(x)$

- This is called bootstrap aggregation/bagging!

- Suppose the ground truth function that we need to predict is $h(x)$

- The prediction of each individual model: $y_b(x) = h(x) + \epsilon_b(x)$

- Error of model $b$: $\epsilon_b(x)$
Bootstrap aggregation

- The average sum-of-squares error for model $b$:
  \[ \mathbb{E}_x[\{y_b(x) - h(x)\}^2] = \mathbb{E}_x[e_b(x)^2] \]

- The average error made by the $b$ models individually:
  \[ E_{AV} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{E}_x[e_b(x)^2] \]

- The expected error of the committee $y_{COM} = \frac{1}{B} \sum_{b=1}^{B} y_b(x)$
  \[ h(x) = \frac{1}{B} \sum_{b=1}^{B} h_b(x) \]
  \[ E_{COM} = \mathbb{E}_x \left[ \left\{ \frac{1}{B} \sum_{b=1}^{B} y_b(x) - h(x) \right\}^2 \right] = \mathbb{E}_x \left[ \left\{ \frac{1}{B} \sum_{b=1}^{B} e_b(x) \right\}^2 \right] \]

- If we assume $\mathbb{E}_x[e_b(x)] = 0$ and $\text{cov}[e_b(x), e_{b'}(x)] = 0$ for $b' \neq b$ then
  \[ \mathbb{E}_x[e_b(x)e_{b'}(x)] = 0 \quad E_{COM} = \frac{1}{B} E_{AV} \]
Bootstrap aggregation

- If we assume \( \mathbb{E}_x[e_b(x)] = 0 \) and \( \mathbb{E}_x[e_b(x)e_{b'}(x)] = 0 \) for \( b' \neq b \)

\[
E_{COM} = \frac{1}{B} E_{AV}
\]

- 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 \( E_{COM} \leq E_{AV} \) 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.

- 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.