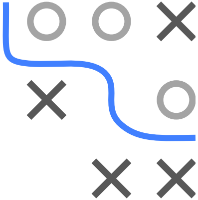
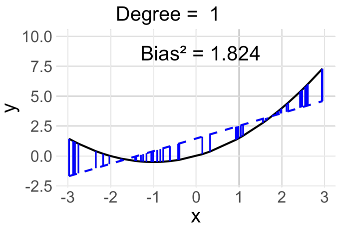
Introduction to Machine Learning

Advanced Risk Minimization
Bias-Variance 1:
Bias-Variance Decomposition

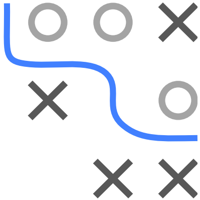




Learning goals

- Decompose GE of learner into
  - bias of learner
  - variance of learner
  - inherent noise of data
- Simulation study demo
- Capacity and overfitting

BIAS-VARIANCE DECOMPOSITION



- Generalization error of learner $\mathcal{I}$: Expected error of model $\mathcal{I}(\mathcal{D}_n) = \hat{f}_{\mathcal{D}_n}$, trained on set of size n, evaled on fresh test sample

$$GE_n(\mathcal{I}) = \mathbb{E}_{\mathcal{D}_n \sim \mathbb{P}_{xy}^n, (\mathbf{x}, y) \sim \mathbb{P}_{xy}} \left[L(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})) \right] = \mathbb{E}_{\mathcal{D}_n, xy} \left[L(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})) \right]$$

- $\mathbb{E}$ taken over all train sets **and** independent test sample. Could also frame this as expected risk (expectation over $\mathcal{D}_n$)

$$GE_n(\mathcal{I}) = \mathbb{E}_{\mathcal{D}_n} \left[\mathbb{E}_{xy} \left[L(y, \hat{f}_{\mathcal{D}_n}(\mathbf{x})) \right] \right] = \mathbb{E}_{\mathcal{D}_n} \left[\mathcal{R}(\hat{f}_{\mathcal{D}_n}) \right]$$

- For L2 loss, can additively decompose $GE_n(\mathcal{I})$ into 3 components
- Assume data is generated by

$$y = f_{\text{true}}(\mathbf{x}) + \epsilon$$

  with 0-mean homoskedastic error $\epsilon \sim (0, \sigma^2)$; independent of $\mathbf{x}$
- Similar decomps exist for other losses expressable as Bregman divergences (e.g. log-loss). One exception is $0/1$ ▸ Brown and Ali 2024

BIAS-VARIANCE DECOMPOSITION

$GE_n(\mathcal{I}) =$

$$\underbrace{\sigma^2}_{\text{Var. of } \epsilon} + \mathbb{E}_x \underbrace{\left[\text{Var}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x})\right]}_{\text{Variance of learner at } \mathbf{x}} + \mathbb{E}_x \underbrace{\left[(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x}))^2\right]}_{\text{Squared bias of learner at } \mathbf{x}}$$

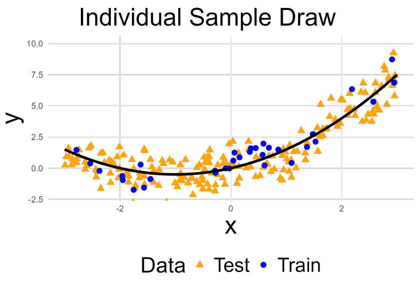
1. First: variance of “pure” **noise** ϵ; aka Bayes, intrinsic or irreducible error; whatever we we do, will never be better
2. Second: how much $\hat{f}_{\mathcal{D}_n}(\mathbf{x})$ **fluctuates** at test $\mathbf{x}$ if we vary training data, averaged over feature space; = learner’s tendency to learn random things irrespective of real signal (overfitting)
3. Third: how “off” are we on average at test locations (underfitting); uses “average model integrated out over all $\mathcal{D}_n$”; models with high capacity have low **bias** and vice versa

SIMULATION EXAMPLE

- DGP with true model:

$$y = x + \frac{x^2}{2} + \epsilon \qquad \epsilon \sim N(0, 1)$$

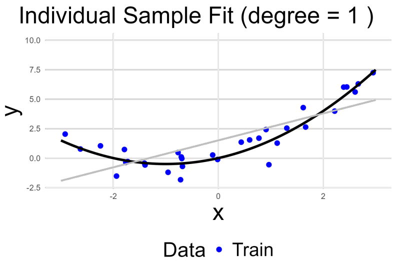
- We will later draw multiple training sets $\mathcal{D}_n$, but only generate one large test set to set to approx. integrate our loss (can nicely do this with simul data)

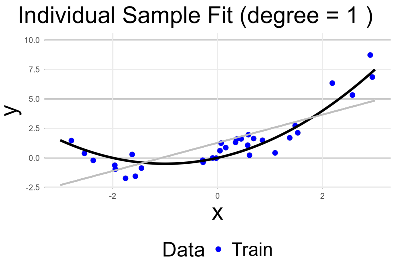


(only part of large test set shown here and in later plots)

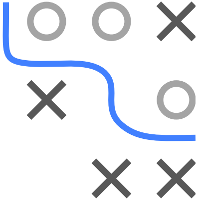
SIMULATION EXAMPLE

- Let's estimate bias and variance by drawing independent data sets from the DGP and averaging
- First, we train several (low capacity) LMs
- These are the $\hat{f}_{\mathcal{D}_n}(\mathbf{x})$, seen as a RV, based on the random data $\mathcal{D}_n$

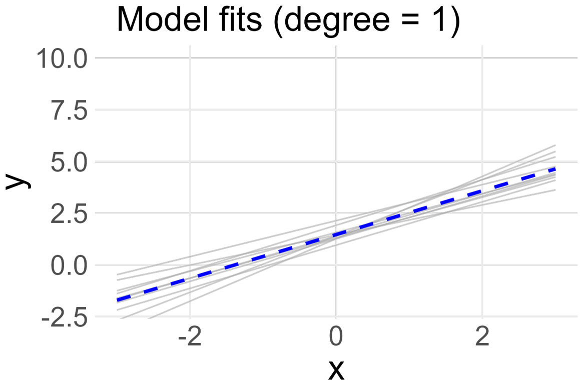




AVERAGE MODEL

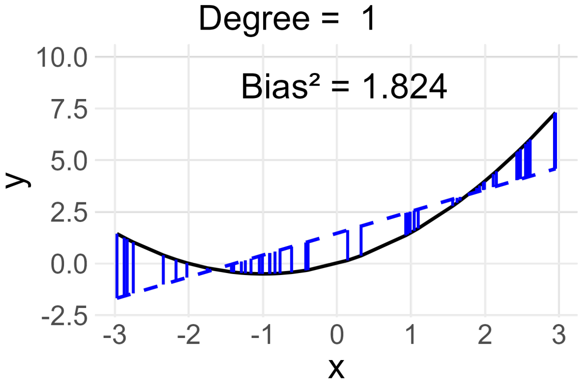


- Average model over different training datasets
- This is $\mathbb{E}_{\mathcal{D}_n}[\hat{f}_{\mathcal{D}_n}(\mathbf{x})]$ in the decomp

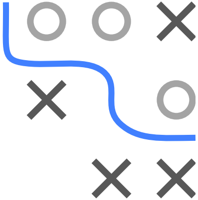


SQUARED BIAS COMPUTATION / ESTIMATION

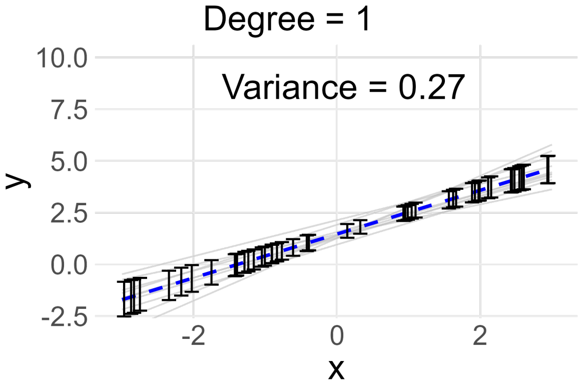
- Compute sq. diff. between avg. and true model at each test x
- Then average over all test points (plot only shows subset)
- This is $\mathbb{E}_x[(f_{\text{true}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x})) \mid \mathbf{x})^2]$



VARIANCE COMPUTATION



- Compute variance of model predictions at each test x
- Then average over all test points (plot only shows subset)
- This is $\mathbb{E}_x[\mathrm{Var}_{\mathcal{D}_n}(\hat{f}_{\mathcal{D}_n}(\mathbf{x}) \mid \mathbf{x})]$

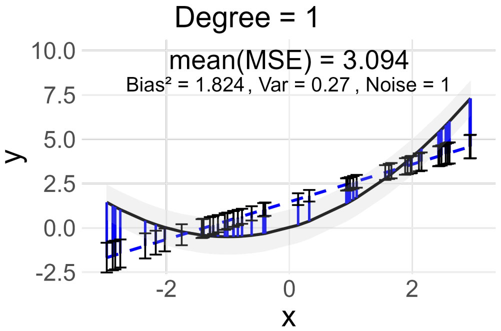


- For irreducible noise component, we know data variance $\sigma^2 = 1$; could also estimate it from residuals

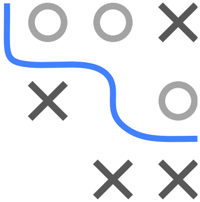
DECOMP RESULT AND COMPARISON WITH MSE

- Decomp result; here bias is largest:

$$GE_n(\mathcal{I}) \approx 1 + 1.824 + 0.270 = 3.094$$

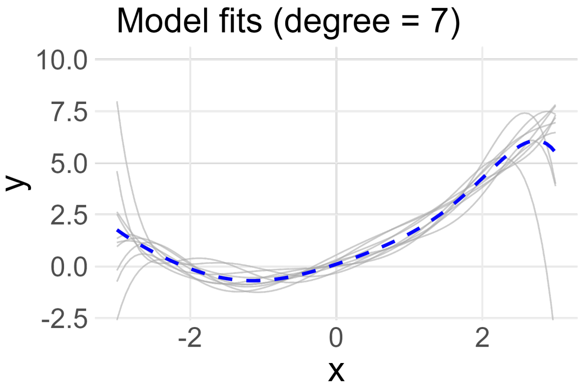


- Regular MSE: For each model, compute MSE on whole test set
- Then we average these MSEs over all models
- Result = 3.094; checks out;
- In general: Error is quite high as we underfitted

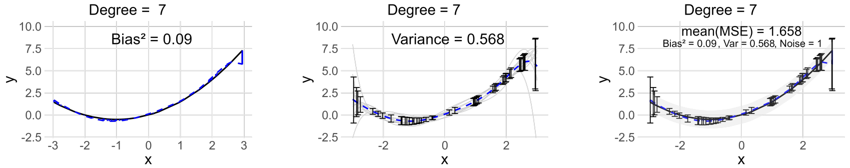


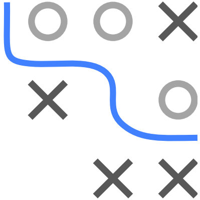
HIGHER COMPLEXITY LEARNER

- Same procedure, but using a high-degree polynomial ($d = 7$). Average model looks good now (low bias)



HIGHER COMPLEXITY LEARNER



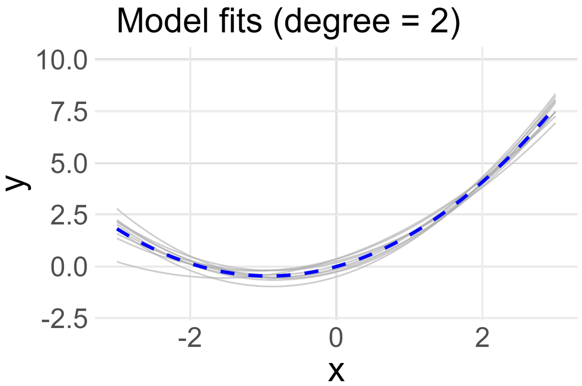


$$GE_n(\mathcal{I}) \approx 1 + 0.09 + 0.568 = 1.658$$

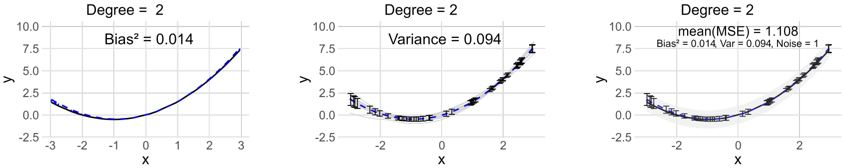
- GE lower than before and hypo space now contains f_{true}
- Bias is much lower, and variance higher
- Higher capacity learner overfits (here).
  We also do not regularize, that would be better
- NB: There is an "edge effect" on LHS, Runge effect,
  leads to some bias as "artifact" here (ignore this)

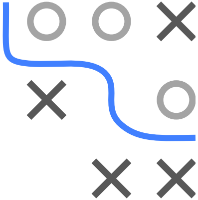
CORRECT COMPLEXITY LEARNER

- What happens if we use a model with the same complexity as the true model (quadratic polynomial)?



CORRECT COMPLEXITY LEARNER



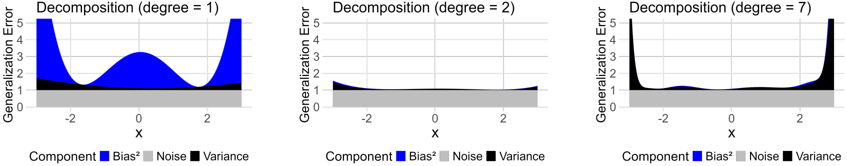


$$GE_n(\mathcal{I}) \approx 1 + 0.014 + 0.094 = 1.108$$

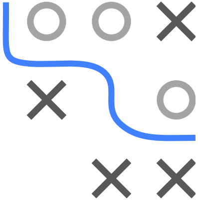
- Naturally: better result
- Lowest bias, low variance
- In any case, variance of the data (irreducible noise, here 1) is a lower bound of GE
- This part remains even when using true model and infinite data

POINT-WISE DECOMPOSITION

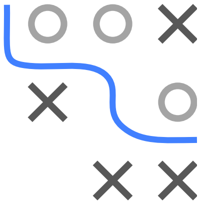
We can also compute these quantities point-wise, showing how each component varies over the domain of *x*



- For LM there is significant bias depending on *x*
- GE for degree 2 is dominated by irreducible noise and model var. at boundaries
- GE for degree 7 is dominated by exploding variance terms near boundaries



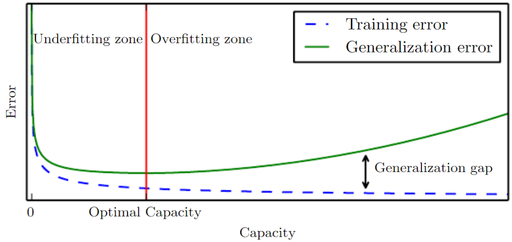
CAPACITY AND OVERFITTING



- Performance of a learner depends on its ability to
  1. **fit** the training data well
  2. **generalize** to new data

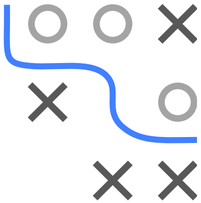
- Failure of the first point is called **underfitting**

- Failure of the second point is called **overfitting**

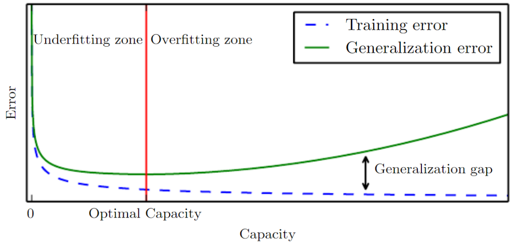


Credit: Ian Goodfellow

CAPACITY AND OVERFITTING



- Tendency of a learner to underfit/overfit is function of its capacity, determined by the type of hypotheses it can learn
- Usually: high capacity $\rightarrow$ low bias $\rightarrow$ better fit on train
- But: high capacity $\rightarrow$ high variance $\rightarrow$ high chance of overfitting
- For such models, regularization (discussed later) is essential
- Even for correctly specified models, generalization error is lower-bounded by irreducible noise σ^2



Credit: Ian Goodfellow