Bias-Variance Decomposition

Machine Learning Course - CS-433
Oct 11, 2023
Nicolas Flammarion
EPFL

Last time

How can we judge if a given predictor is good?

How to select the best models of a family?

→ Bound the difference between the true and empirical risks

→ Split data into train and test sets (learn with the train and test on the test)

Motivation: Hyperparameters search (which often control the complexity)

But we haven't investigated the role of the complexity of the class

Today

How does the risk behave as a function of the complexity of the model class?

⇒ Bias-Variance tradeoff

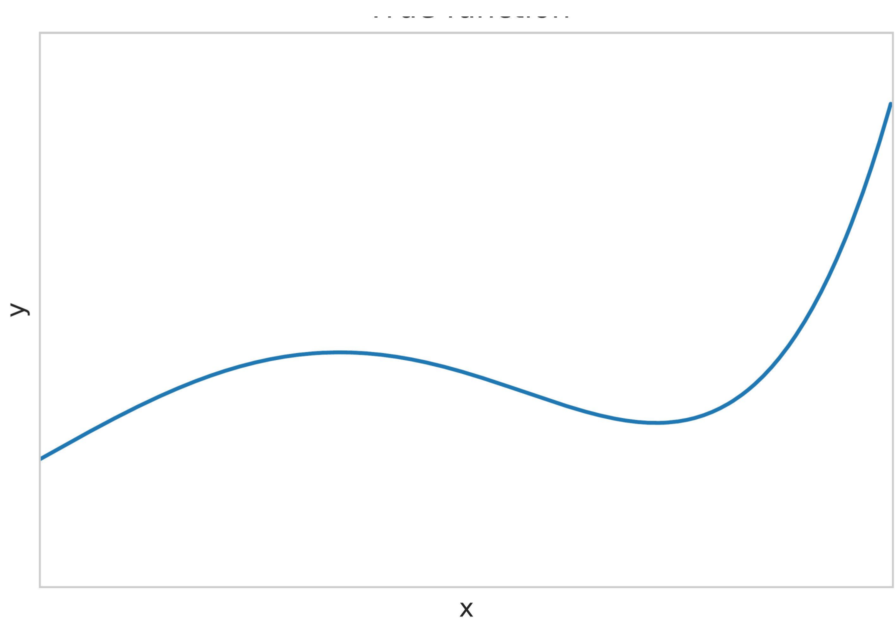
It will help us to decide how complex and rich we should make our model

Before: quantitative

Now: qualitative

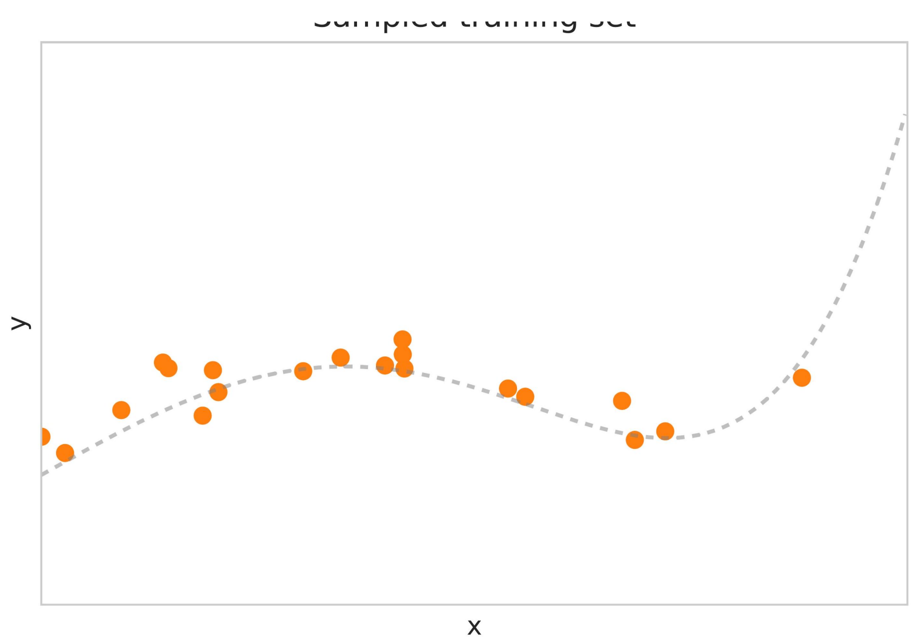
A small experiment: 1D-regression

True function



A small experiment: 1D-regression

Sampled training set

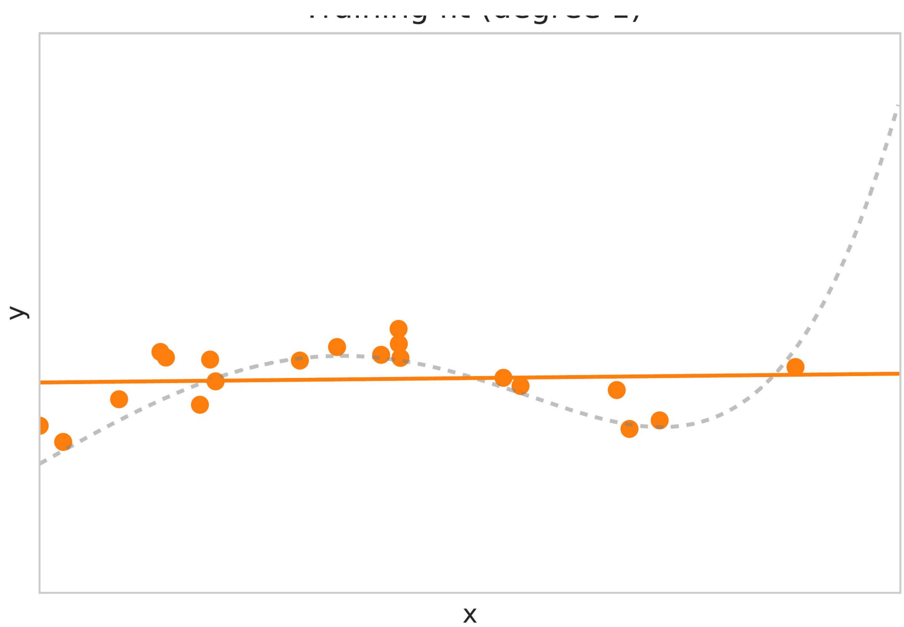


Linear regression using polynomial feature expansion $\left(x, x^2, x^3, \cdots, x^d\right)$ The

maximum degree d measures the complexity of the class
$\Rightarrow$ How far should you go?

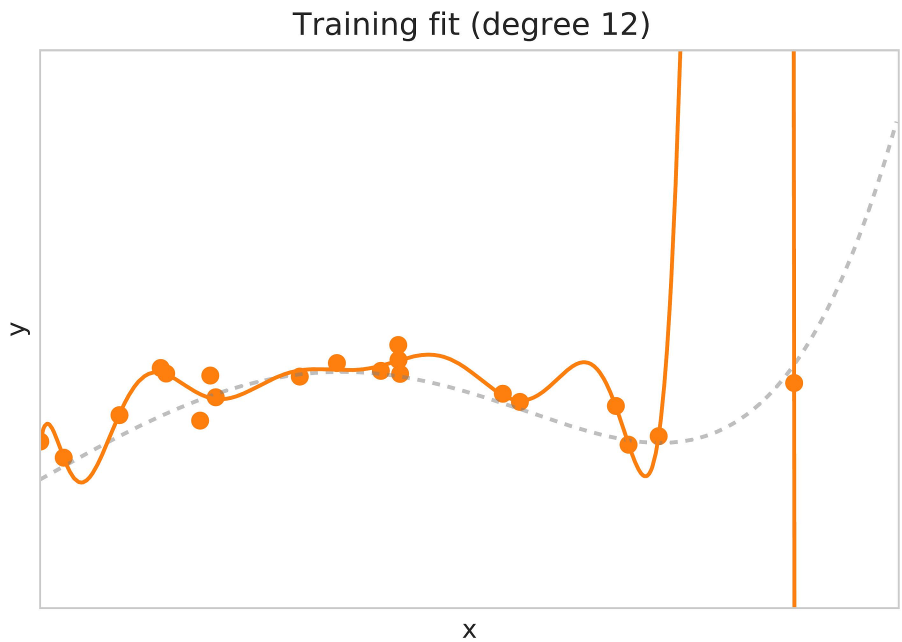
Simple model: bad fit

Training fit (degree 1)



No linear function would be a good predictor. The model class is not rich enough

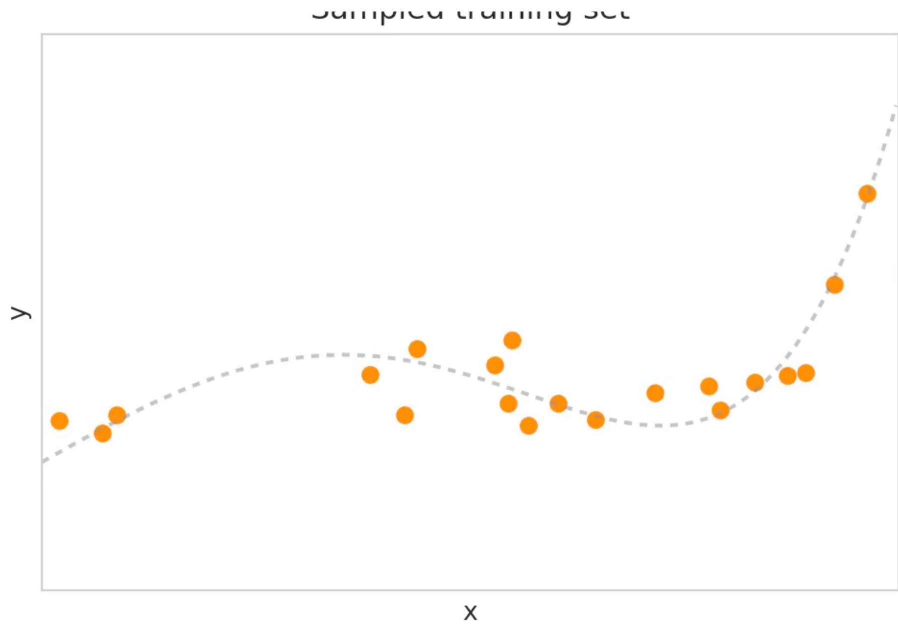
Complex model: good fit?



High degree polynomial will be a good fit. But?

But there is randomness in the data

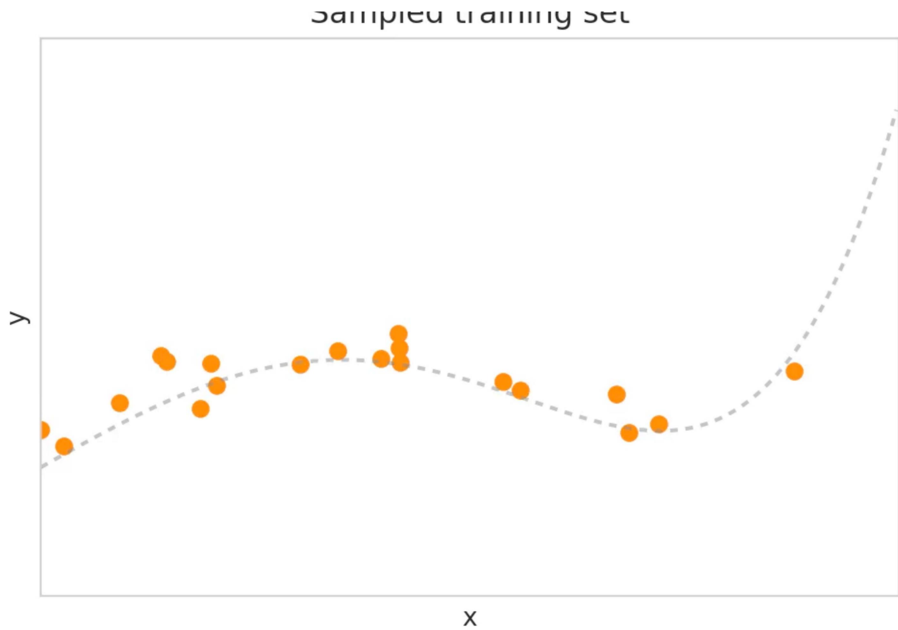
Sampled training set



We have observed one particular S_{train} but we could have observed several others!

But there is randomness in the data

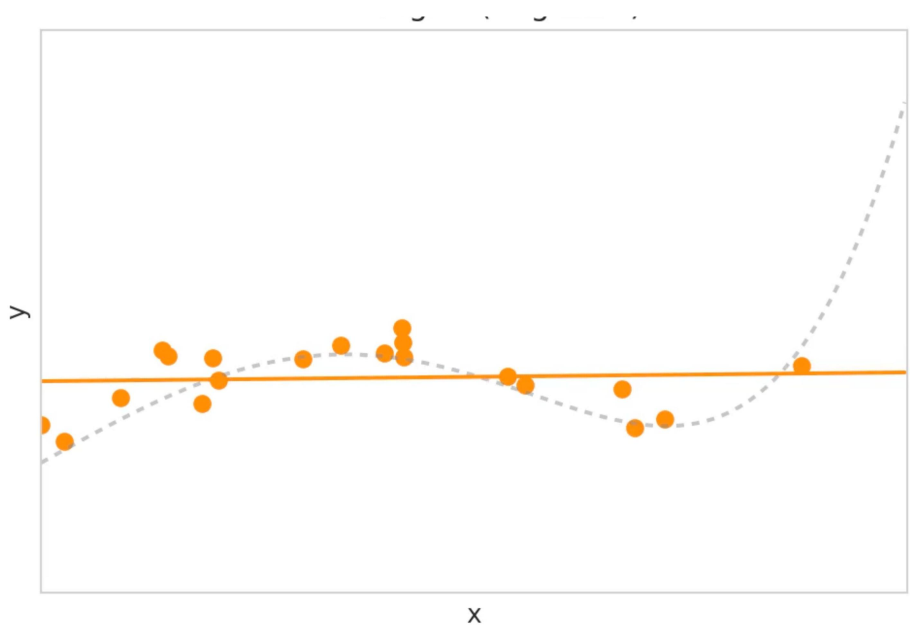
Sampled training set



Even if we keep the same $(x_1, \cdots, x_n)$, we have variability in the observed $(y_1, \cdots, y_n)$

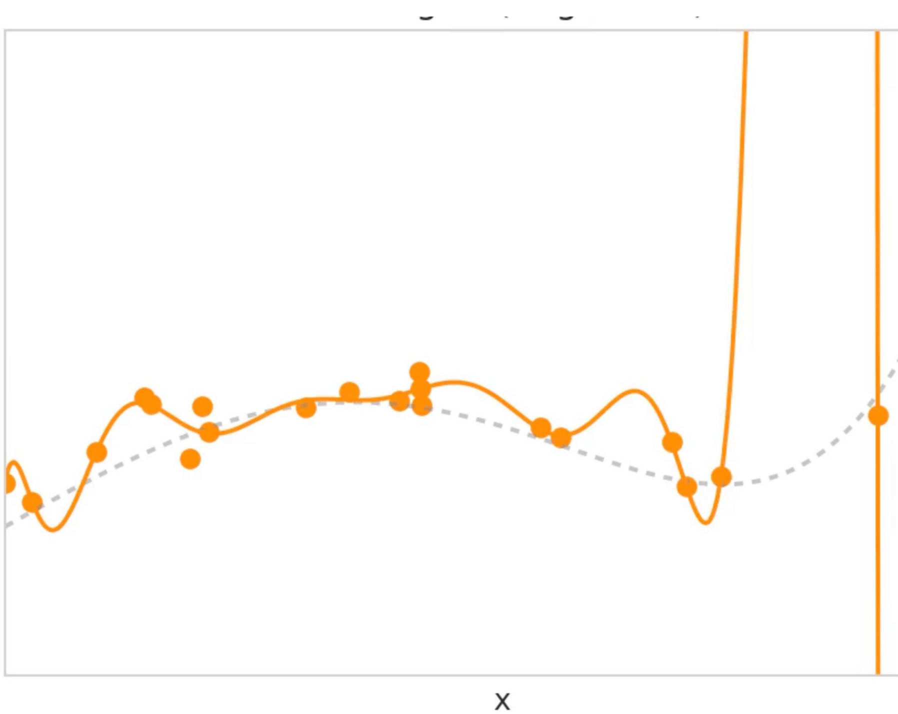
Thus there is randomness in the predictions

Training fit (degree 1)



Moving a single observation will cause only a small shift in the position of the line

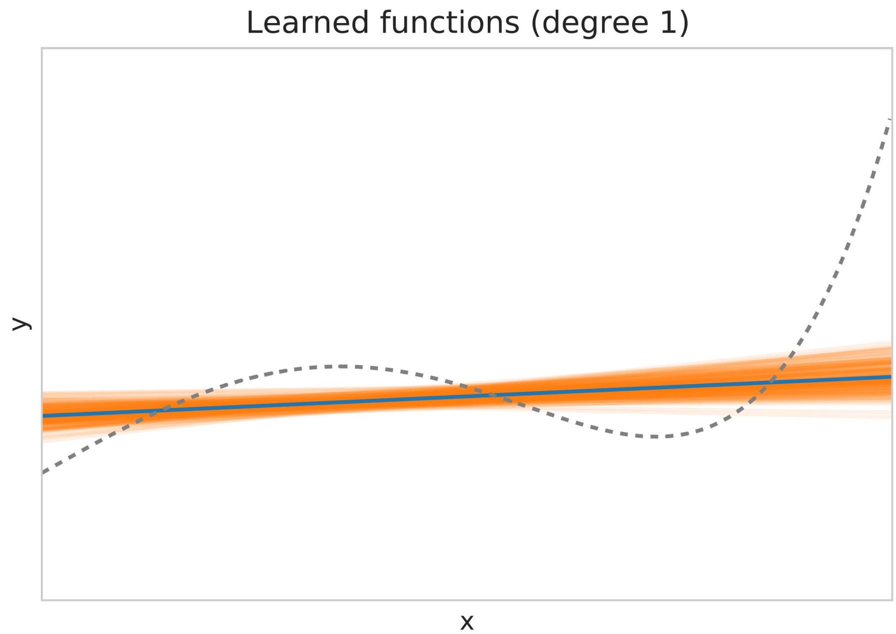
Underfitting Training fit (degree 12)



Changing one of the observations may change the prediction considerably

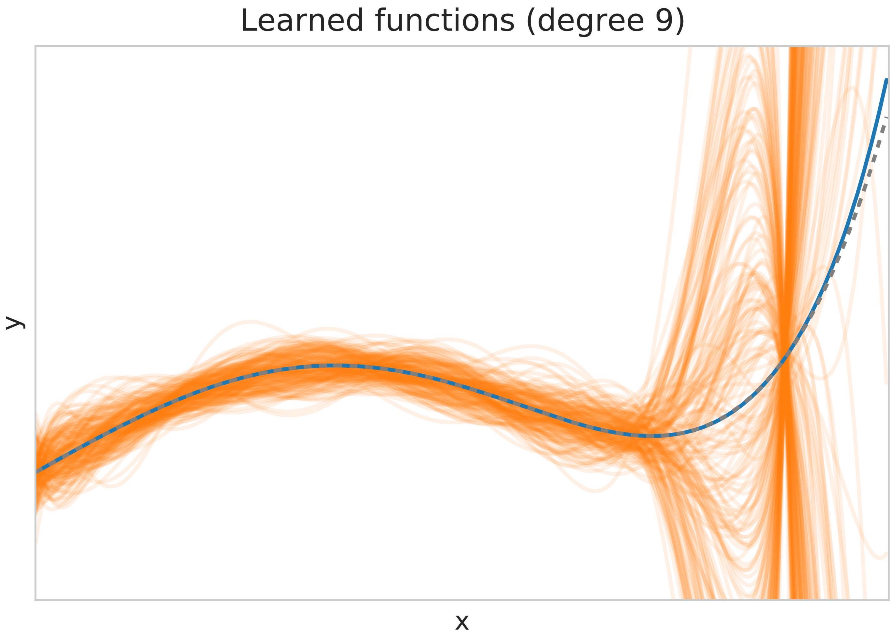
Overfitting
Simple models are less sensitive

Simple models have large bias but low variance



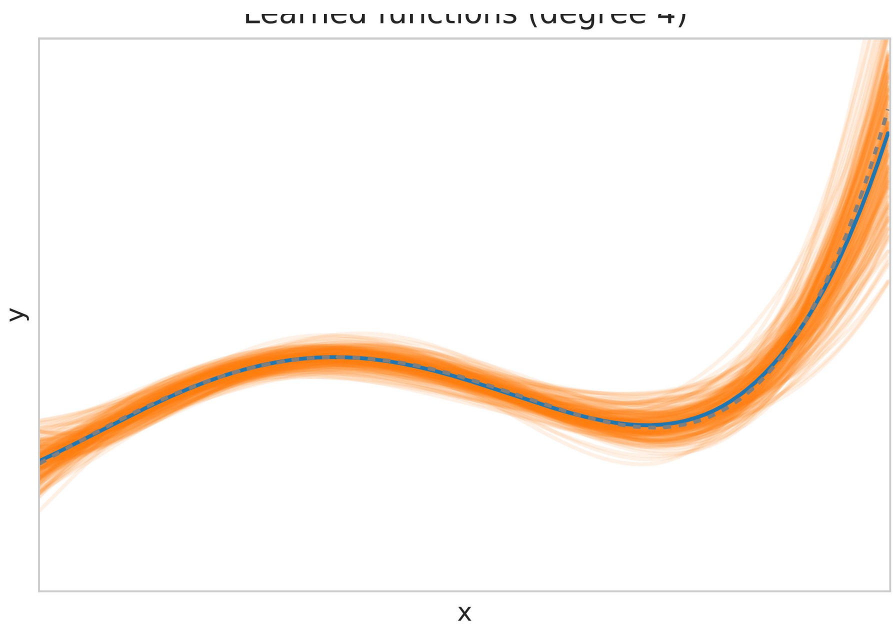
The average of the predictions f_S does not fit well the data: large bias The variance of the predictions f_S as a function of S is small: small variance

Complex models have low bias but high variance



The average of the predictions f_S fits well the data: small bias
The variance of the predictions f_S as a function of S is large: large variance

We need to balance bias & variance correctly



Data model: output perturbed by some noise

Output
True model perturbed by some noise
Joint distribution $(x, y) \sim \mathscr{D}$

$$y = f(x) + \varepsilon$$

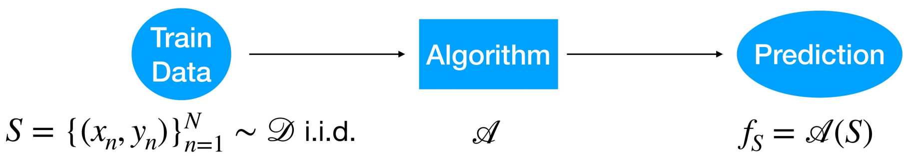
True model
f arbitrary and unknown function
The model is **not realizable**, i.e.,
f is not in our model class

Input
$x \sim \mathscr{D}_x$
(fixed but unknown)

Noise
$\varepsilon \sim \mathscr{D}_\varepsilon$ i.i.d.,
independent of x
$\mathbb{E}[\varepsilon] = 0$

We consider the square loss and will provide a decomposition of the true error

Error Decomposition

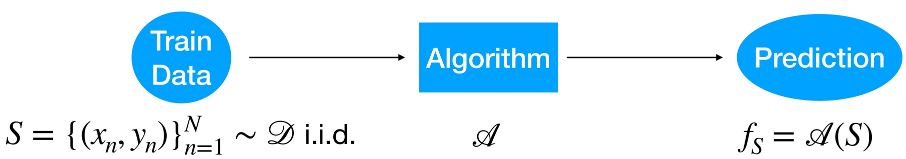


We are interested in how the expected error of f_S :

$$\mathbb{E}_{(x,y)\sim D}\left[(y - f_S(x))^2\right]$$

behaves as a function of the train set S and model class complexity

Error Decomposition

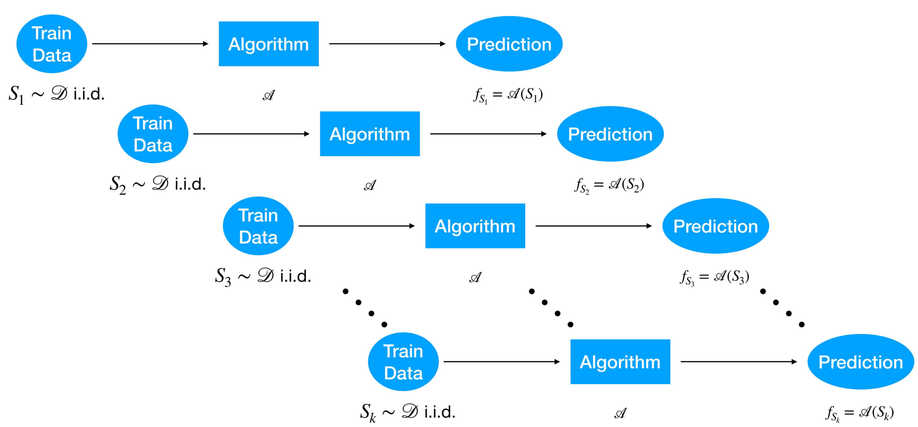


The decomposition will hold true at every single point x. Therefore, to simplify, we consider the expected error of f_S for a fixed element x_0 :

$$L\left(f_S\right) = \mathbb{E}_{\varepsilon\sim\mathscr{D}_\varepsilon}\left[\left(f\left(x_0\right) + \varepsilon - f_S\left(x_0\right)\right)^2\right]$$

This is a random variable. The randomness comes for the train set S

We run the experiment many times



We are interested in the average and the variance of the predictions $(f_{S_1}, \cdots, f_{S_k})$ over these multiple runs

A decomposition in three terms

We are interested in the expectation of the true risk over the training set S

$$\begin{aligned}\mathbb{E}_{S\sim\mathscr{D}}\left[L\left(f_S\right)\right] &= \mathbb{E}_{S\sim\mathscr{D}}\left[\mathbb{E}_{\varepsilon\sim D_\varepsilon}\left[\left(f\left(x_0\right)+\varepsilon-f_S\left(x_0\right)\right)^2\right]\right] \\ &= \mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\left(f\left(x_0\right)+\varepsilon-f_S\left(x_0\right)\right)^2\right]\end{aligned}$$

We will decompose this quantity in three non-negative terms and will interpret each of these terms

First we expand the square:

$$\begin{aligned}&\mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\left(f\left(x_0\right)+\varepsilon-f_S\left(x_0\right)\right)^2\right] = \mathbb{E}_{\varepsilon\sim\mathscr{D}_\varepsilon}\left[\varepsilon^2\right] \\ &+ 2\mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\varepsilon\left(f\left(x_0\right)-f_S\left(x_0\right)\right)\right] \\ &+ \mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-f_S\left(x_0\right)\right)^2\right]\end{aligned}$$

Using that $\mathbb{E}_{\varepsilon\sim\mathscr{D}}[\varepsilon]=0$ and $\varepsilon \perp\!\!\!\perp S$:

- $\mathbb{E}_{\varepsilon\sim\mathscr{D}_\varepsilon}\left[\varepsilon^2\right] = \mathrm{Var}_{\varepsilon\sim\mathscr{D}_\varepsilon}[\varepsilon]$
- $\mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\varepsilon\left(f\left(x_0\right)-f_S\left(x_0\right)\right)\right] = \mathbb{E}_{\varepsilon\sim\mathscr{D}\varepsilon}[\varepsilon] \times \mathbb{E}_{S\sim\mathscr{D}}\left[f\left(x_0\right)-f_S\left(x_0\right)\right] = 0$

Therefore
$\mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\left(f\left(x_0\right)+\varepsilon-f_S\left(x_0\right)\right)^2\right] = \mathrm{Var}_{\varepsilon\sim D_\varepsilon}[\varepsilon]+\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-f_S\left(x_0\right)\right)^2\right]$

Trick: we add and subtract the constant term $\mathbb{E}_{S'\sim D}\left[f_{S'}\left(x_0\right)\right]$, where S' is a second training set independent from S

$$\begin{aligned}\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-f_S\left(x_0\right)\right)^2\right] =&\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]+\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)^2\right] \\ =&\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathcal{D}}\left[f_{S'}\left(x_0\right)\right]\right)^2+\left(\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)^2\right. \\ &\left.+2\left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)\left(\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)\right]\end{aligned}$$

Cross-term:

$$\begin{aligned}&\mathbb{E}_{S\sim\mathscr{D}}[\left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)\cdot\left(\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)] \\ &\quad= \left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)\cdot\mathbb{E}_{S\sim\mathscr{D}}\left[\left(\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)\right] \\ &\quad= \left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)\cdot\left(\mathbb{E}_{S'\sim\mathcal{D}}\left[f_{S'}\left(x_0\right)\right]-\mathbb{E}_{S\sim\mathscr{D}}\left[f_S\left(x_0\right)\right]\right)=0.\end{aligned}$$

$$\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f\left(x_0\right)-f_S\left(x_0\right)\right)^2\right] = \left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)^2+\mathbb{E}_{S\sim\mathscr{D}}\left[\left(\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]-f_S\left(x_0\right)\right)^2\right]$$

Bias-Variance Decomposition

We obtain the following decomposition into three positive terms:

$$\mathbb{E}_{S\sim\mathscr{D},\varepsilon\sim\mathscr{D}_\varepsilon}\left[\left(f\left(x_0\right)+\varepsilon-f_S\left(x_0\right)\right)^2\right] = \mathrm{Var}_{\varepsilon\sim\mathscr{D}_\varepsilon}[\varepsilon] \longleftarrow \text{Noise variance}$$

$$\text{Bias } \longrightarrow + \left(f\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)^2$$

$$\text{Variance } \longrightarrow + \mathbb{E}_{S\sim\mathscr{D}}\left[\left(f_S\left(x_0\right)-\mathbb{E}_{S'\sim\mathscr{D}}\left[f_{S'}\left(x_0\right)\right]\right)^2\right]$$

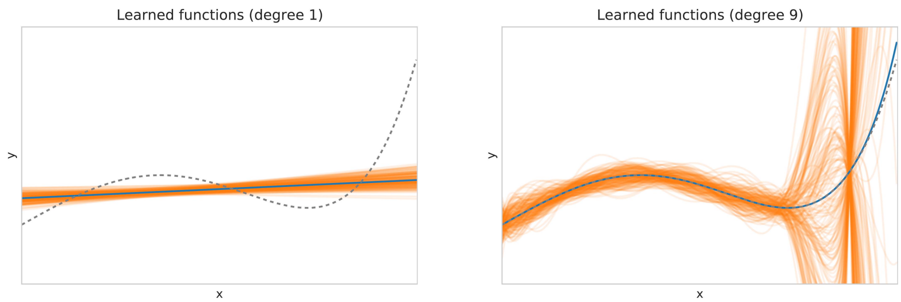
each of which always provides a lower bound of the true error

$\Rightarrow$ To minimize the true error, we must choose a method that achieves low bias and low variance simultaneously

Noise: a strict lower bound on the achievable error

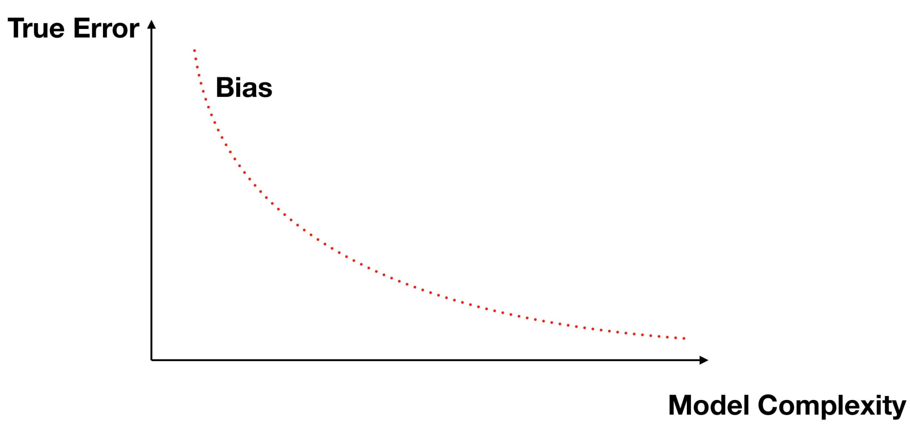
- It is not possible to go below the noise level
- Even if we know the true model f, we still suffer from the noise: $L(f) = \mathbb{E}\left[\varepsilon^2\right]$
- It is not possible to predict the noise from the data since they are independent

Bias: $\left(f\left(x_0\right)-\mathbb{E}_{S\sim\mathscr{D}}\left[f_S\left(x_0\right)\right]\right)^2$



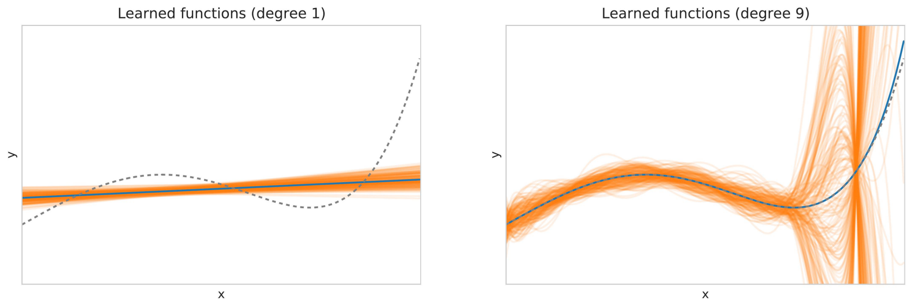
- Squared of the difference between the actual value $f\left(x_0\right)$ and the expected prediction
- It measures how far off in general the models' predictions are from the correct value
- If model complexity is low, bias is typically high
- If model complexity is high, bias Is typically low

Bias: $\left(f\left(x_0\right)-\mathbb{E}_{S\sim\mathscr{D}}\left[f_S\left(x_0\right)\right]\right)^2$



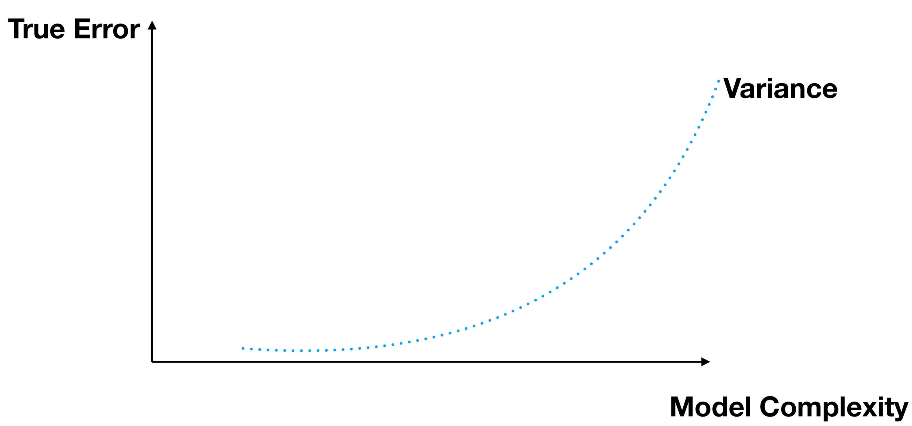
- Squared of the difference between the actual value $f\left(x_0\right)$ and the expected prediction
- It measures how far off in general the models' predictions are from the correct value
- If model complexity is low, bias is typically high
- If model complexity is high, bias is typically low

Variance: $\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f_S\left(x_0\right)-\mathbb{E}_{S\sim\mathscr{D}}\left[f_S\left(x_0\right)\right]\right)^2\right]$



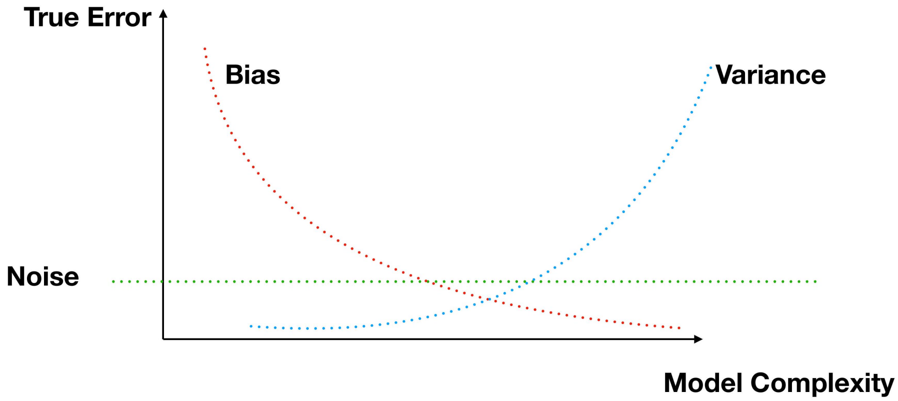
- Variance of the prediction function
- It measures the variability of predictions at a given point across different training set realizations
- If we consider complex models, small variations in the training set can lead to significant changes in the predictions

Variance: $\mathbb{E}_{S\sim\mathscr{D}}\left[\left(f_S\left(x_0\right)-\mathbb{E}_{S\sim D}\left[f_S\left(x_0\right)\right]\right)^2\right]$



- Variance of the prediction function
- It measures the variability of predictions at a given point across different training set realizations
- If we consider complex models, small variations in the training set can lead to significant changes in the predictions

Bias Variance tradeoff and U-shape curve



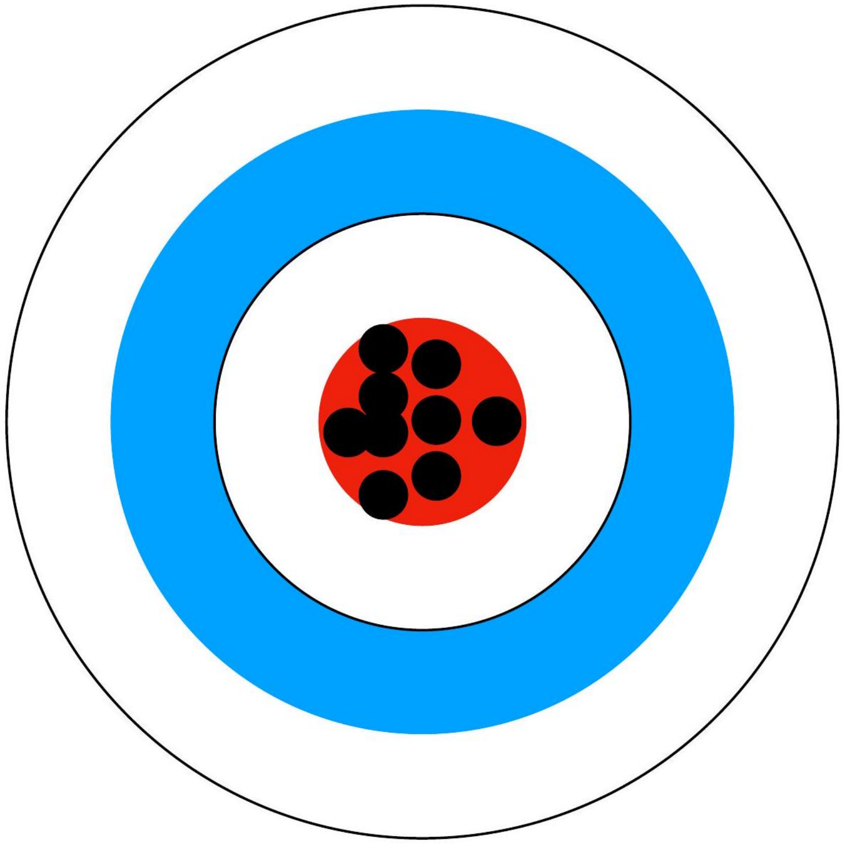
- If model complexity is too low, approximation will be poor (underfitting)
- If model complexity is too high, it may cause issues with variance (overfitting)

$\Rightarrow$ This phenomenon is known as the bias-variance tradeoff

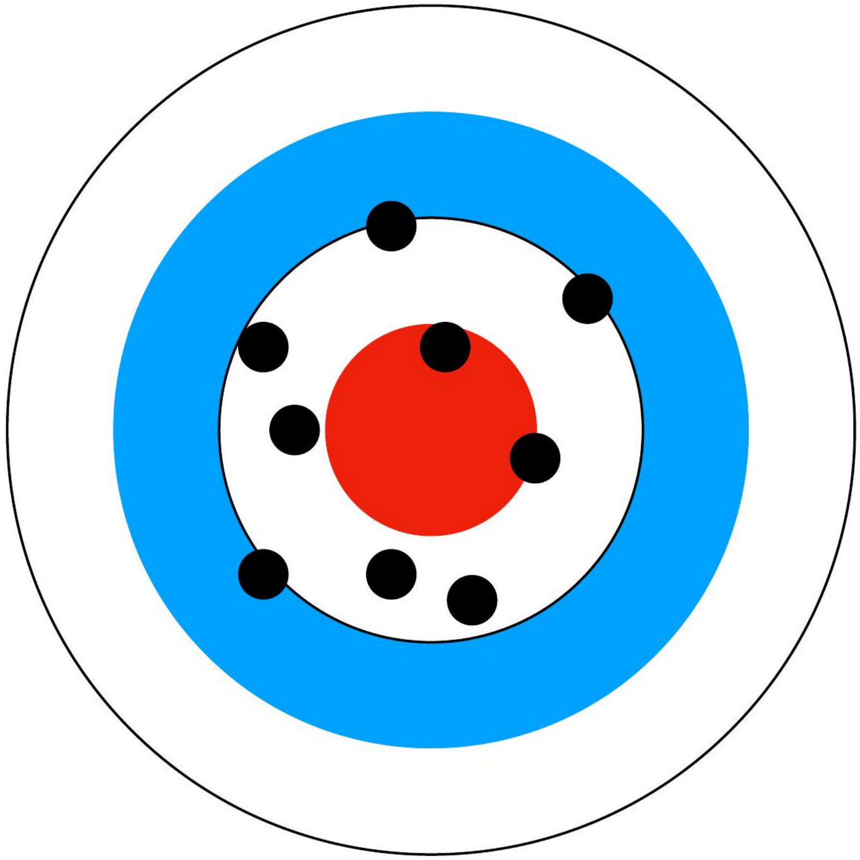
Challenge: Identify a method that ensures both low variance and low bias

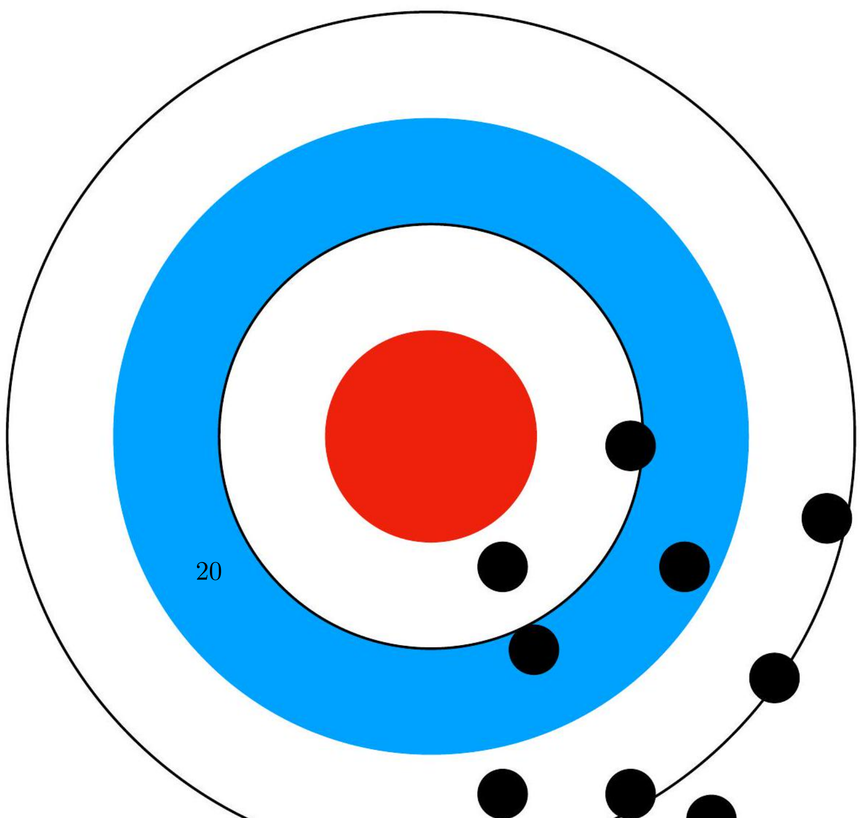
Conclusion

Low Variance
Low Bias



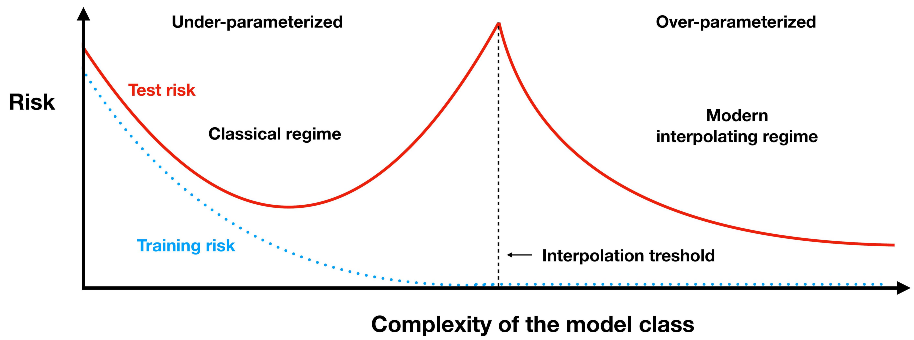
High Bias





But this depends on the algorithm!

Double descent curve



Reconciling modern machine-learning practice and the classical bias-variance trade-off Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal, PNAS, 2019