MODEL SELECTION

Itay Evron

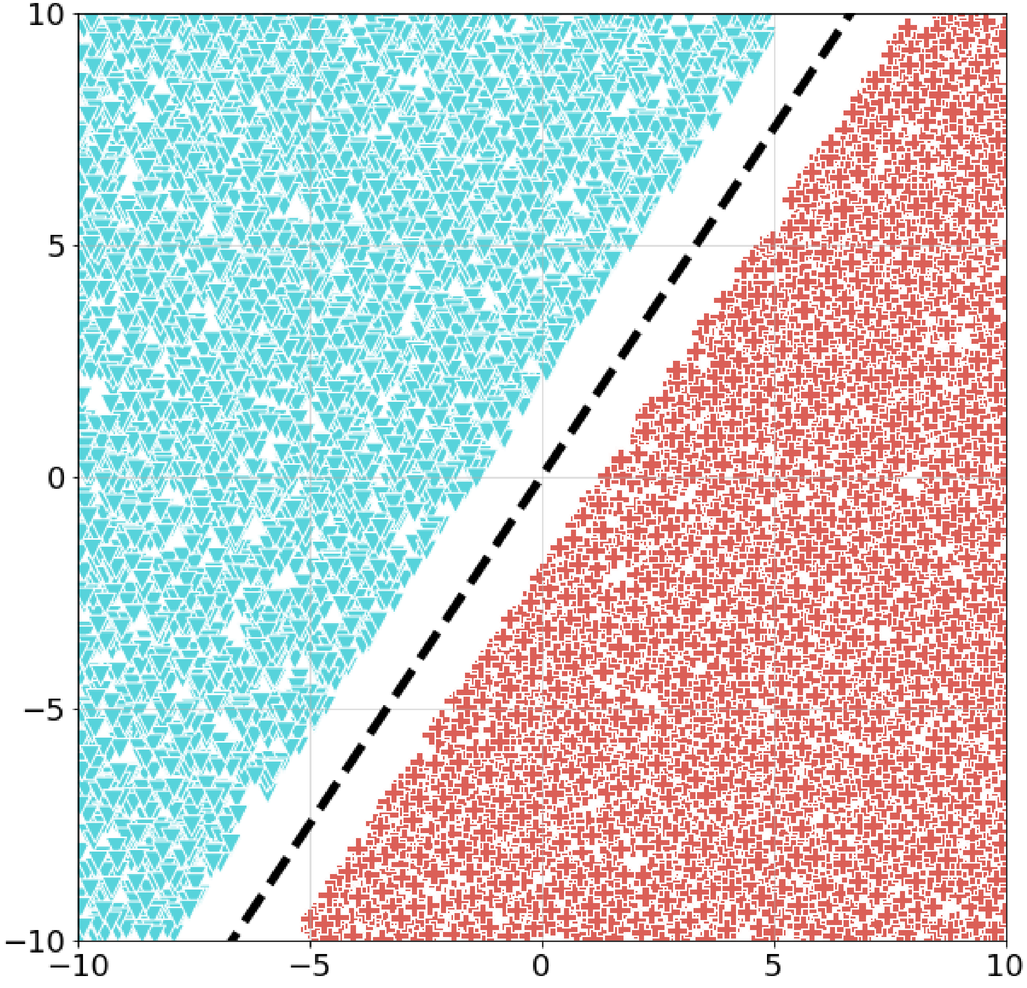
Outline

- Today's tutorial is different: more empirical than analytical
- Recap
  - Bias-variance error decomposition
  - Bias-variance tradeoff
- Demo I: Separable data
- Demo II: Inseparable data
- Model selection

DEMO I:
SEPARABLE DATA

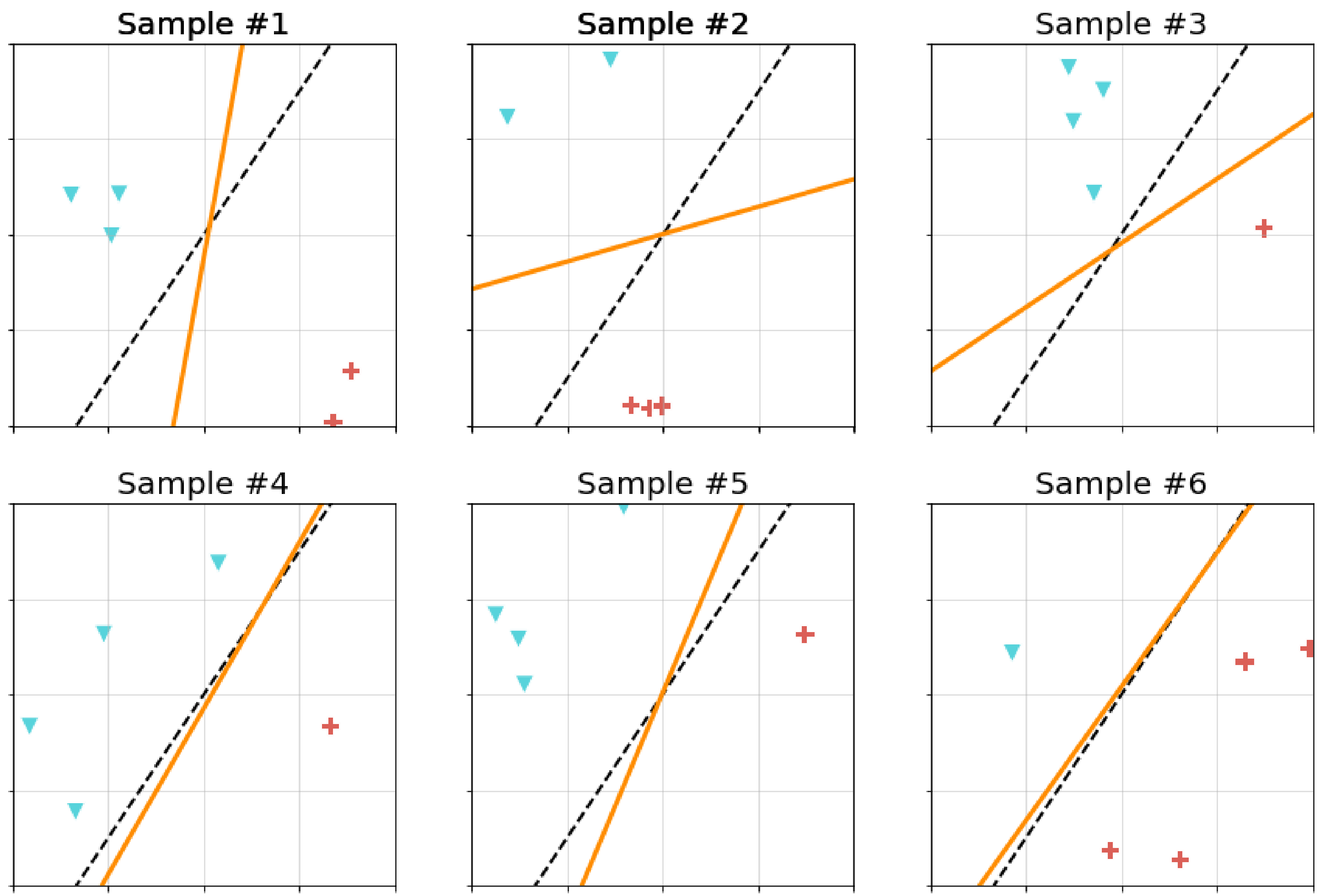
Linearly separable data

Max-margin separator learned on all data

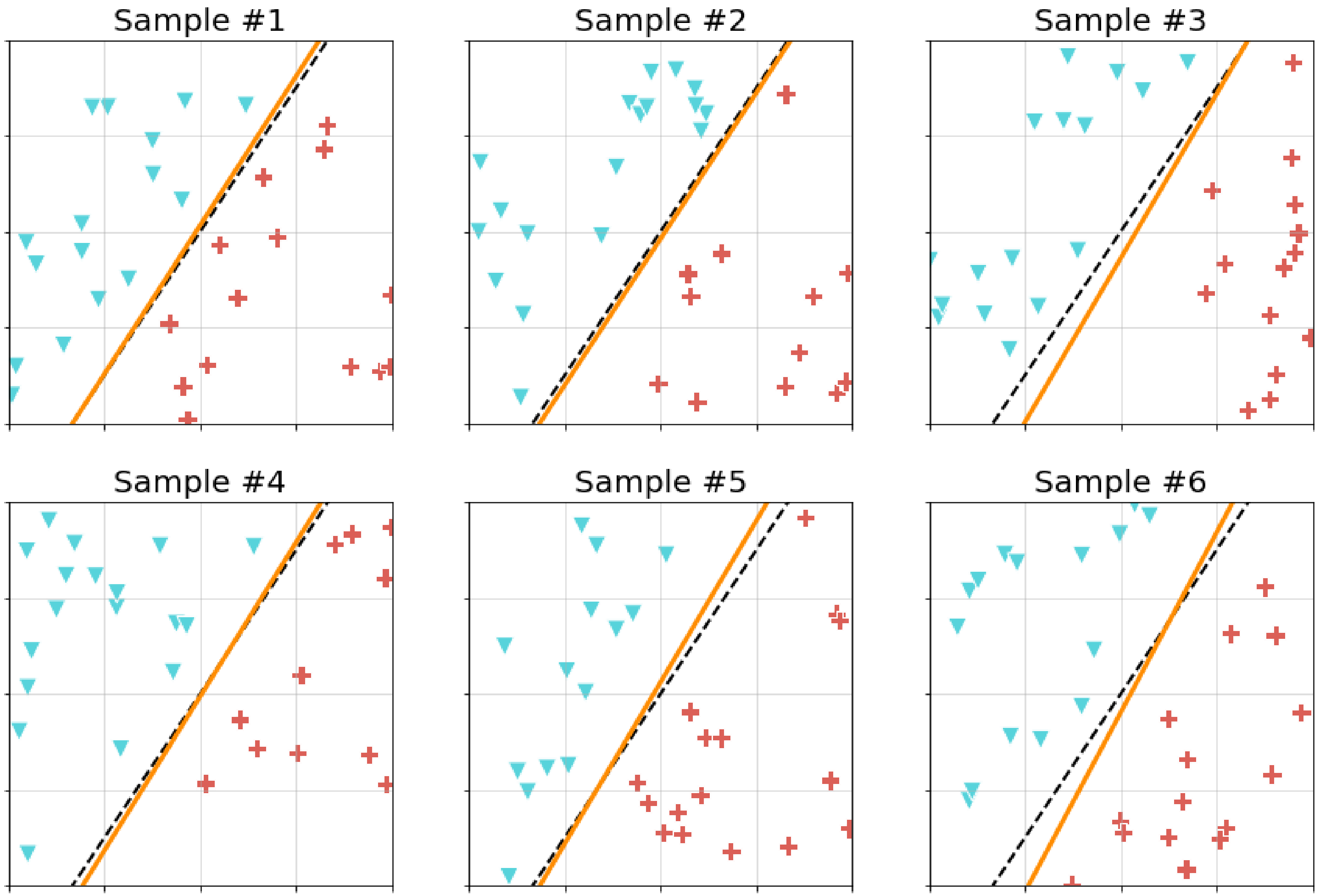


That's the entire distribution!

Random samples when $m = 5$

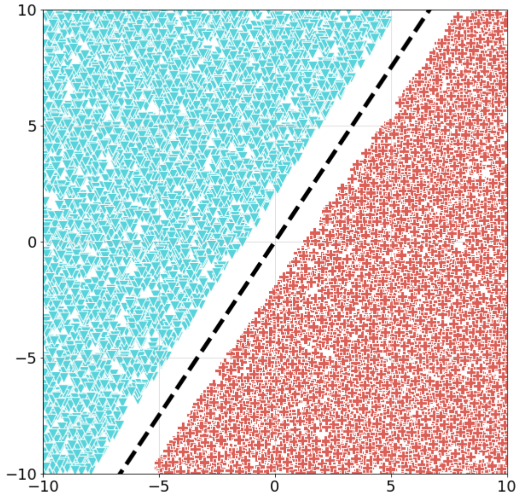


Random samples when $m = 30$

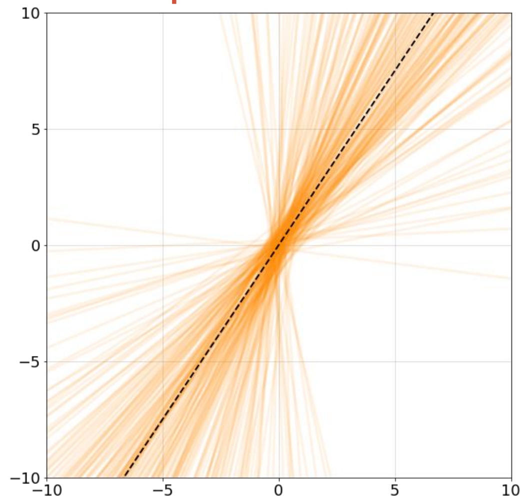


More samples $\Longrightarrow$ lower variance

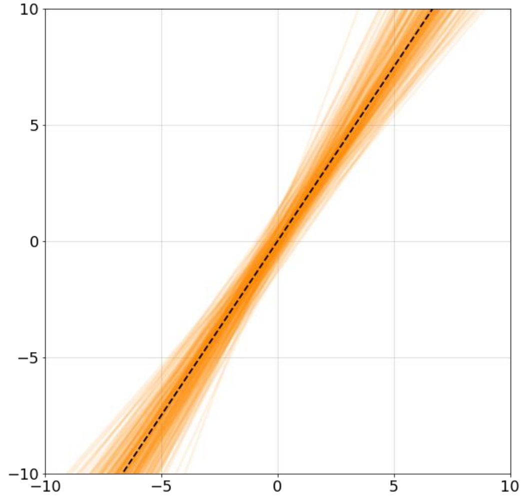
All data



Sample size: $m = 5$



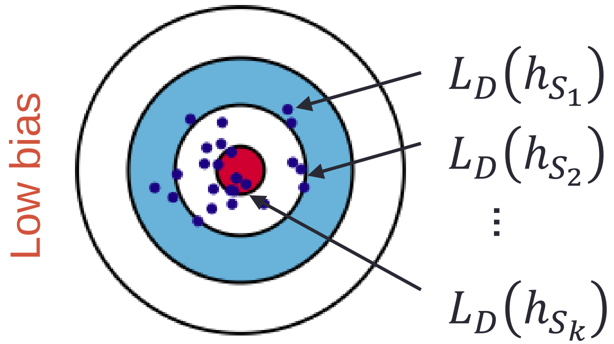
Sample size: $m = 30$



Everything is
a random variable!

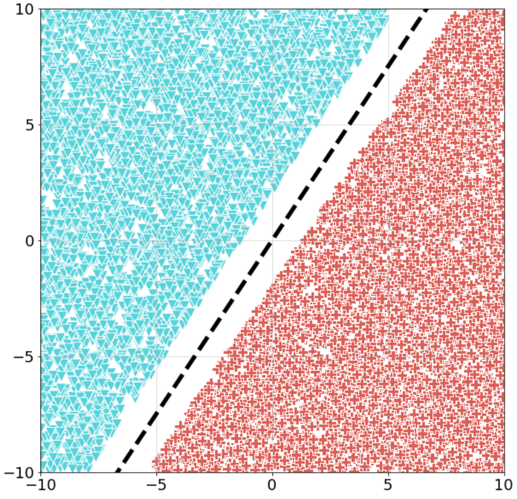
- The sample S_i
- The hypothesis h_{S_i}
- The loss $L_D(h_{S_i})$

High variance

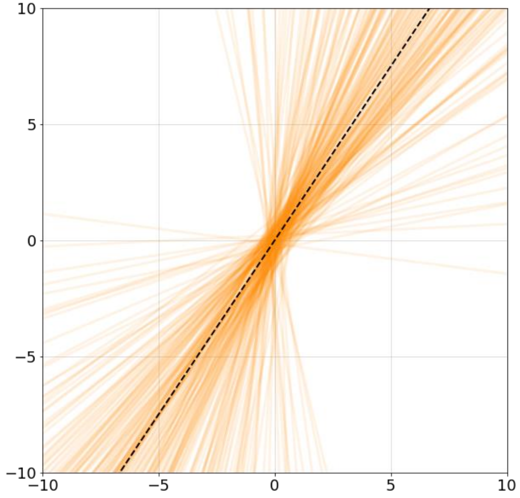


More samples $\Longrightarrow$ lower variance

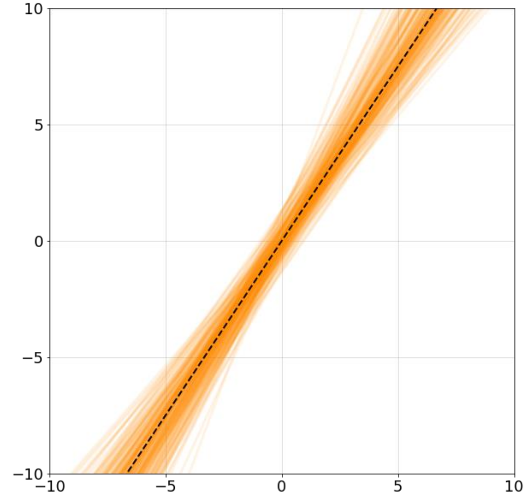
All data



Sample size: $m = 5$



Sample size: $m = 30$

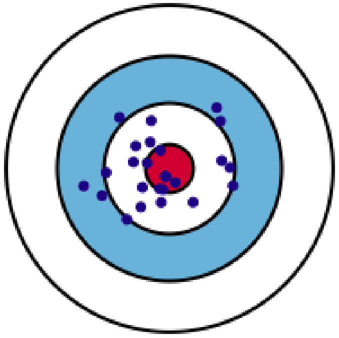


Everything is
a random variable!

- The sample S_i
- The hypothesis h_{S_i}
- The loss $L_D(h_{S_i})$

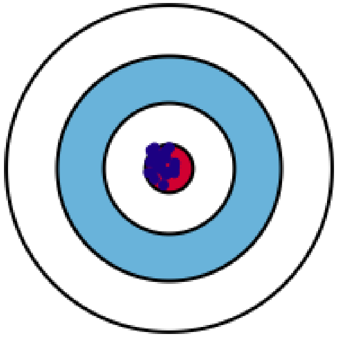
High variance

Low bias



Low variance

Low bias



BIAS-VARIANCE ERROR DECOMPOSITION

Recap: Bias-variance error decomposition

- Three interpretable sources of error:

$$\mathbb{E}_{S\sim D^m}\left[L_D^{sqr}(h_S)\right] = \mathbb{E}_x\left[\left(\bar{h}(x) - \bar{y}(x)\right)^2\right] + \mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right] + \mathbb{E}_{x,y}\left[(\bar{y}(x) - y)^2\right]$$

expected error ***bias²*** ***variance*** ***noise***

Recap: Bias-variance error decomposition

- Three interpretable sources of error:

$$\underbrace{\mathbb{E}_{S\sim D^m}\left[L_D^{sqr}(h_S)\right]}_{\textit{expected error}} = \underbrace{\mathbb{E}_x\left[\left(\bar{h}(x) - \bar{y}(x)\right)^2\right]}_{\textit{bias}^2} + \underbrace{\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]}_{\textit{variance}} + \underbrace{\mathbb{E}_{x,y}\left[(\bar{y}(x) - y)^2\right]}_{\textit{noise}}$$

- **Noise:**
  - Property of data distribution (i.e., the statistical relation between x and y)
  - In this tutorial: assume "realizability", i.e., $\exists f, \forall x$: $\underbrace{y = f(x)}_{\text{deterministic}}$, i.e., **no noise**!

Recap: Bias-variance error decomposition

- (without noise) ~~Three~~ Two interpretable sources of error:

$$\mathbb{E}_{S\sim D^m}\left[L_D^{sqr}(h_S)\right] = \mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right] + \mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right] + \mathbb{E}_{x,y}\left[\left(\bar{y}(x) - y\right)^2\right]$$

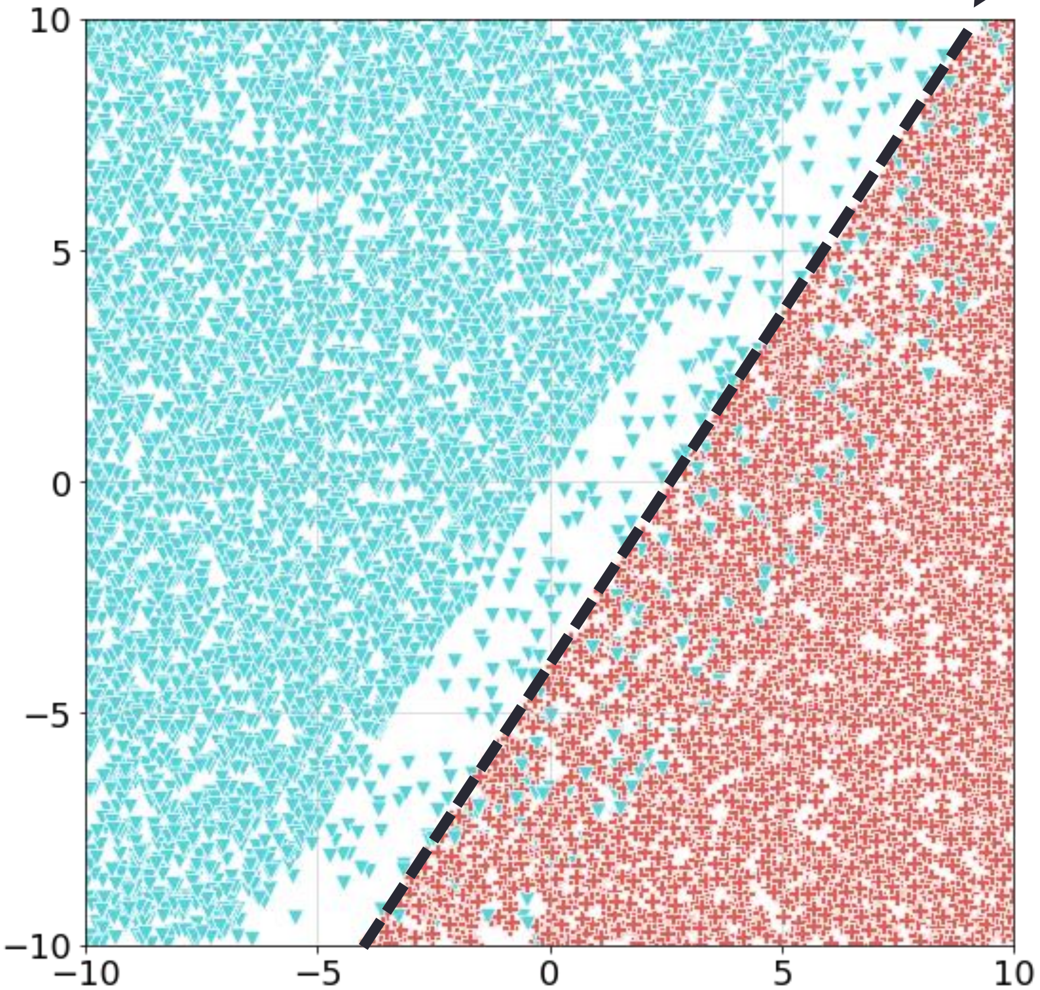
expected error ***bias²*** ***variance*** *noise*

- We will understand these quantities in the following slides.

DEMO II:
INSEPARABLE DATA

Linearly inseparable data

(linear) Separator with lowest generalization error



That's the entire distribution!

Recap: Soft SVM

- Data is <u>not</u> linearly separable; hence we use Soft SVM
- Two conflicting objectives:

$$\underset{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \ \lambda\|\boldsymbol{w}\|_2^2 + \frac{1}{m}\sum_{i\in[m]} \max\{0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\}$$

$$\underset{\boldsymbol{w}\in\mathbb{R}^d, b\in\mathbb{R}}{\operatorname{argmin}} \ \|\boldsymbol{w}\|_2^2 + C\sum_{i\in[m]} \max\{0, 1 - y_i(\boldsymbol{w}^\top \boldsymbol{x}_i + b)\}$$

Complexity penalty (Regularization)
Prefer "simpler" models

Margin violation penalty
How much the i^th example *violates* the margin

Recap: Kernel SVM

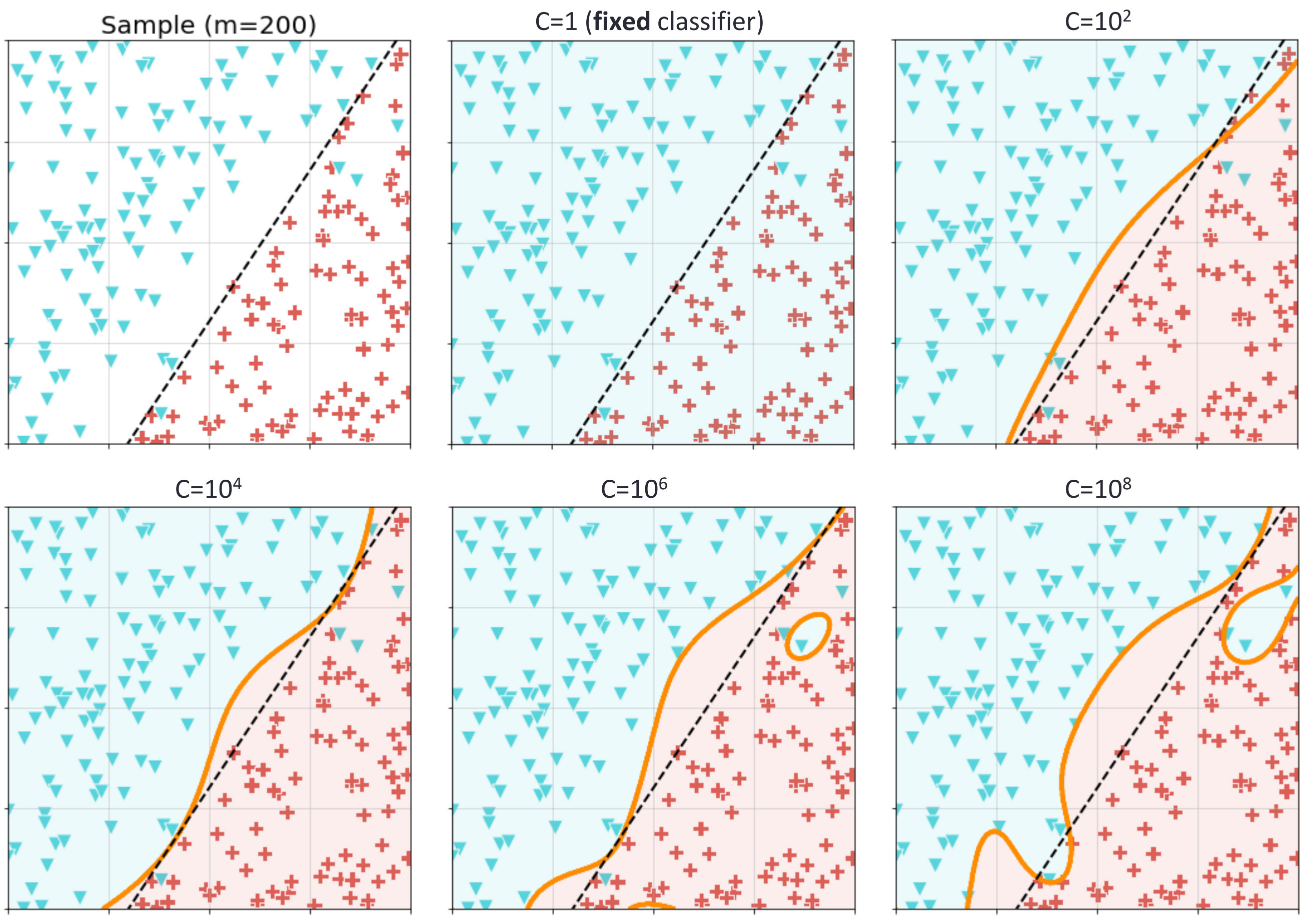
- To make things more interesting, we use an RBF kernel
- Solve an optimization problem equivalent to

$$\underset{\boldsymbol{w}\in\mathbb{R}^{d'},b\in\mathbb{R}}{\operatorname{argmin}} \|\boldsymbol{w}\|_2^2 + C\sum_{i\in[m]} \max\{0, 1 - y_i(\boldsymbol{w}^\top \phi(\boldsymbol{x}_i) + b)\}$$

RBF feature mapping

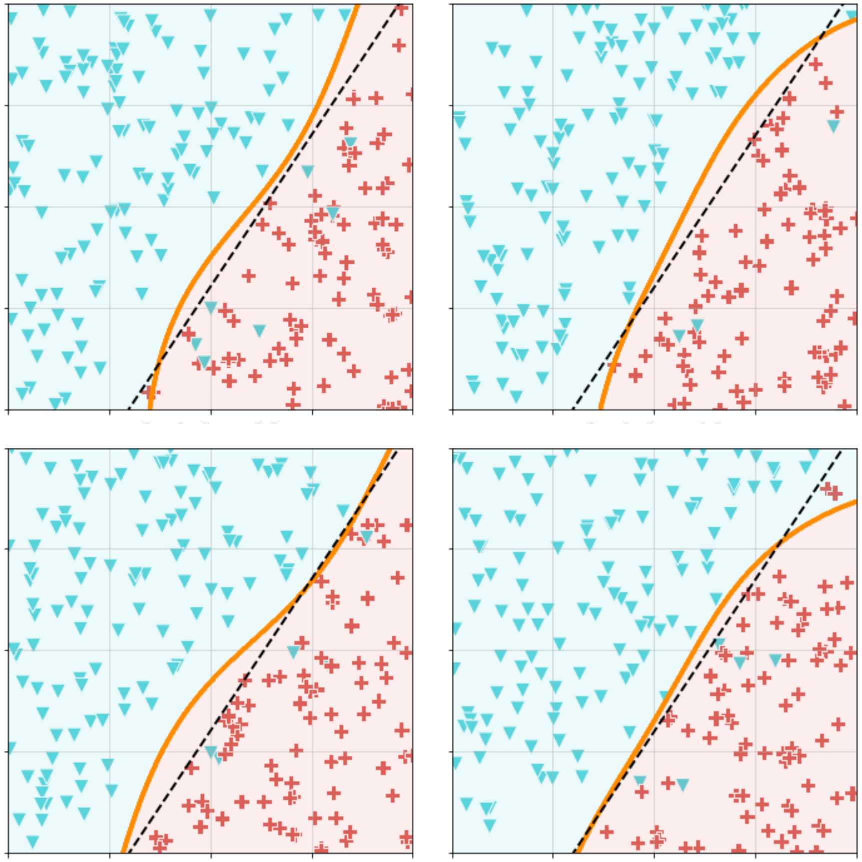
Larger $C \implies$ more complex models

$$\underset{\boldsymbol{w}\in\mathbb{R}^{d'},b\in\mathbb{R}}{\operatorname{argmin}} \|\boldsymbol{w}\|_2^2 + C\sum_{i\in[m]} \max\{0, 1 - y_i(\boldsymbol{w}^\top\phi(\boldsymbol{x}_i) + b)\}$$

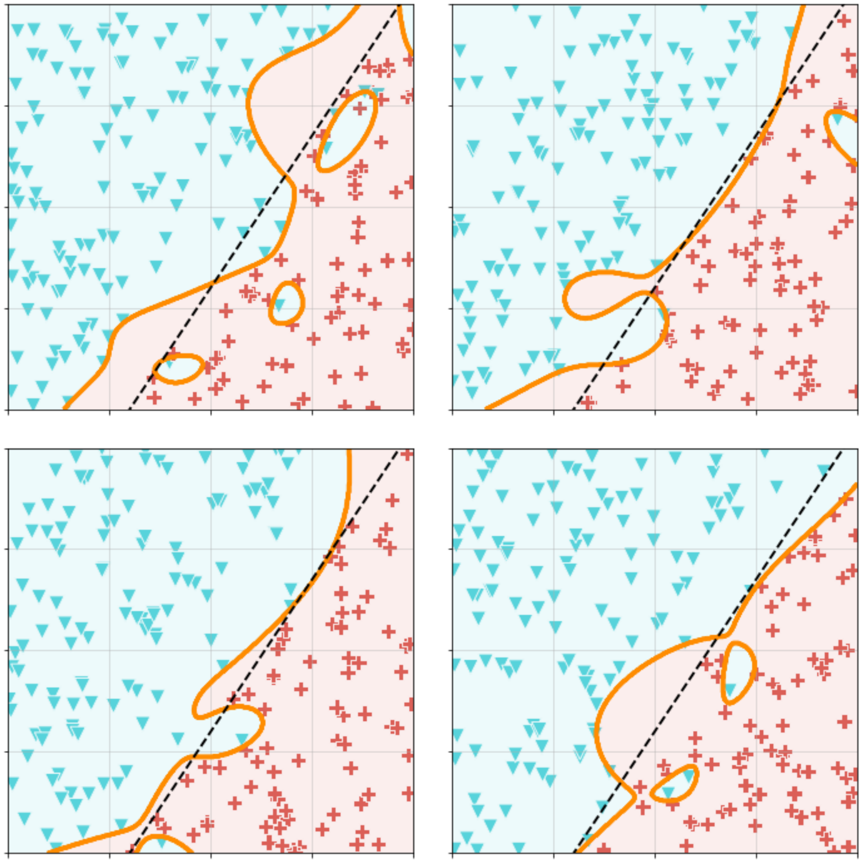


More complex models ⟹ higher variance

Low complexity ($C=10^2$)

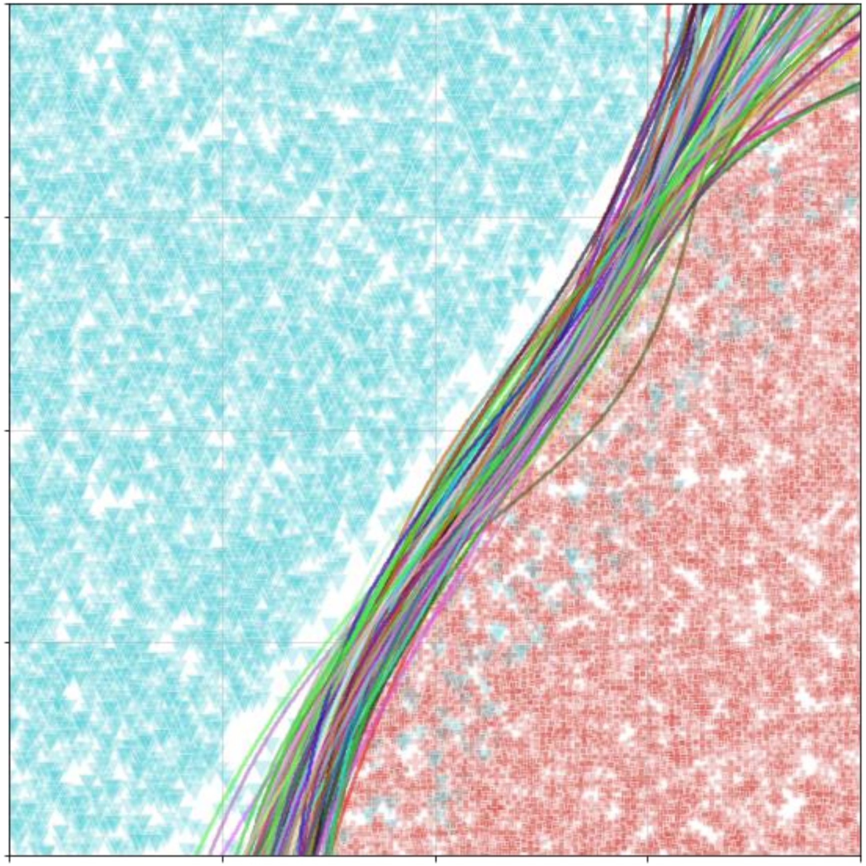


High complexity ($C=10^8$)

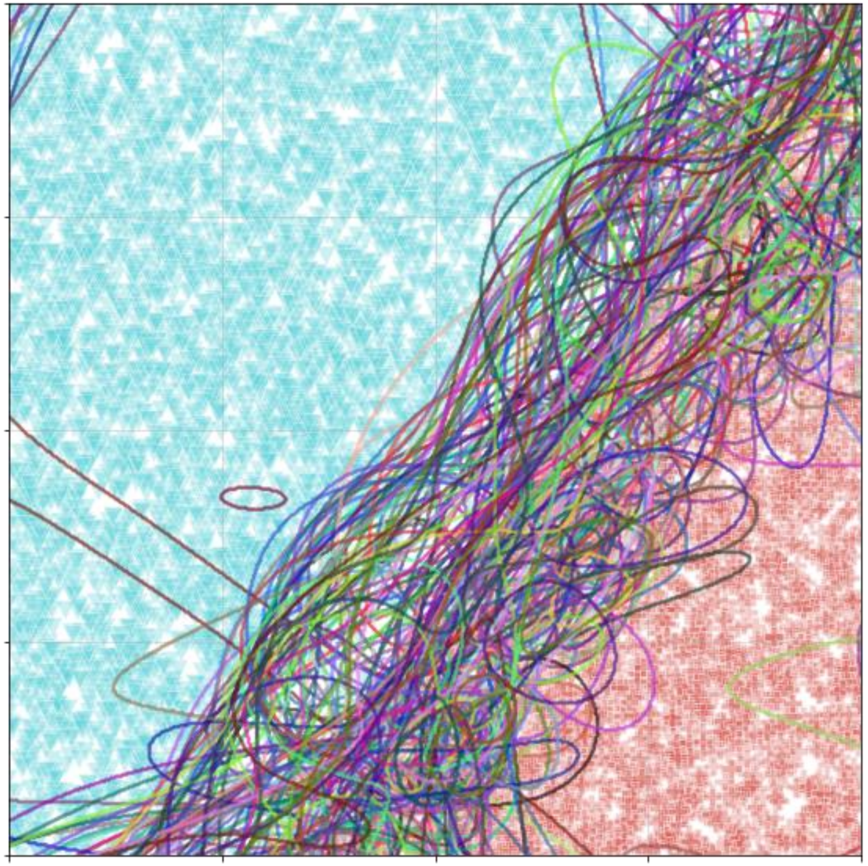


More complex models ⟹ higher variance

Low complexity ($C=10^2$)



High complexity ($C=10^8$)

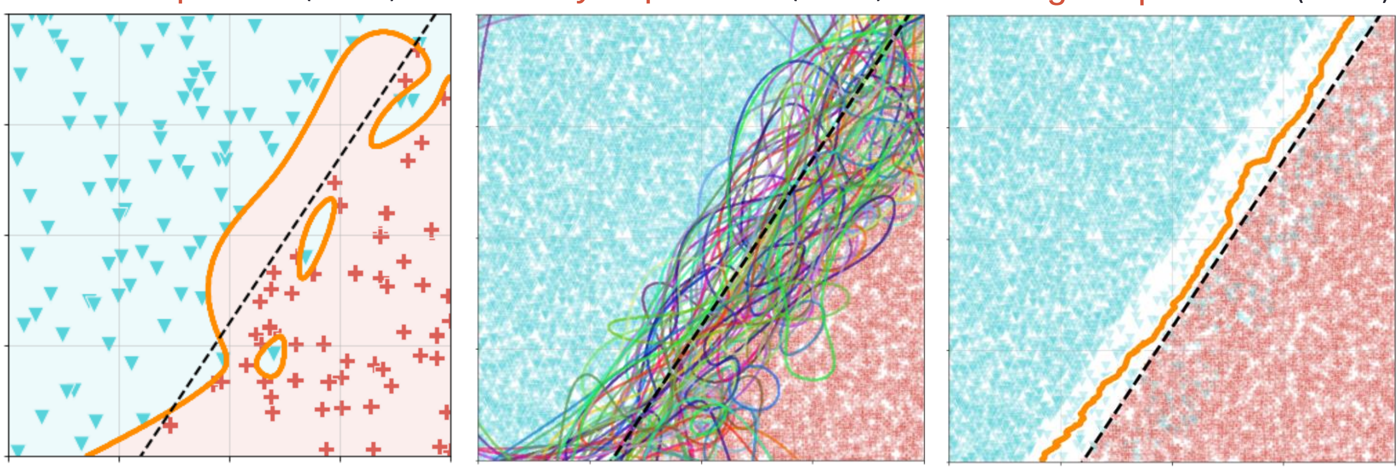


Formalizing the variance

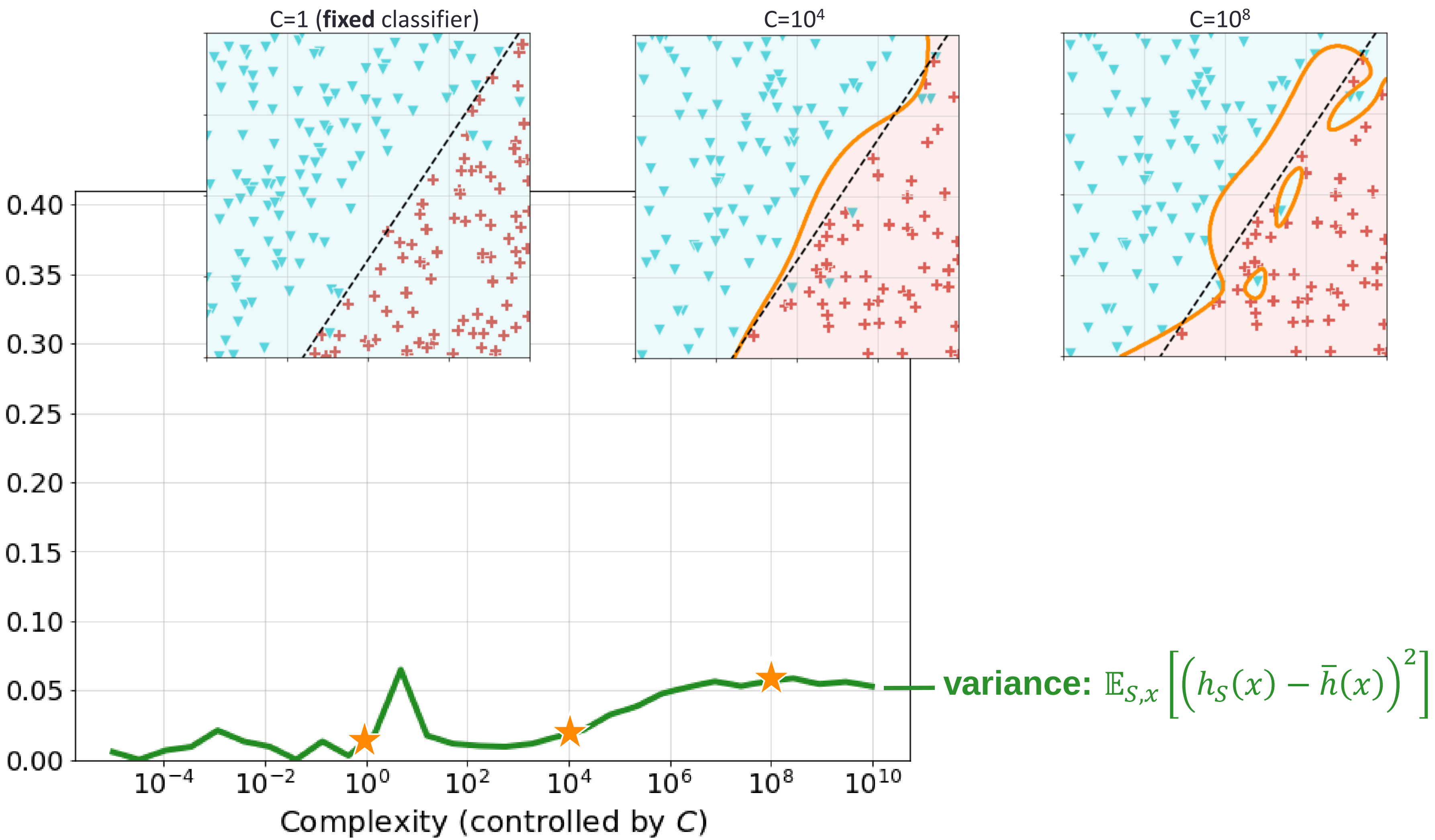
- **Variance:** (of algorithm; w.r.t. S)
  - Measures how output hypotheses h_S vary

    (how “sensitive” the learning algorithm is to changes in its input S)
  - Formally defined as: $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$
  - Average hypothesis $\bar{h}$ as reference point (asks: relative to $\bar{h}$, how specialized is h_S to S?)
- **The “average” hypothesis:** $\bar{h} = \mathbb{E}_{S\sim D^m}[h_S]$

One separator (C=10^8) Many separators (C=10^8) Average separator $\bar{h}$ (C=10^8)

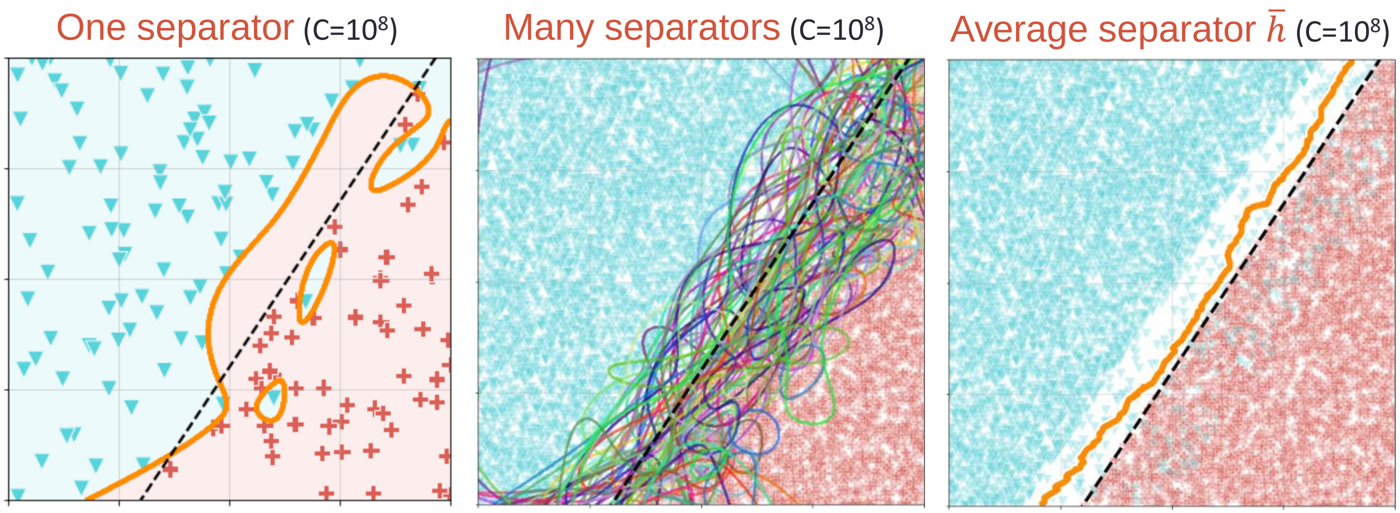


More complex models ⟹ higher variance

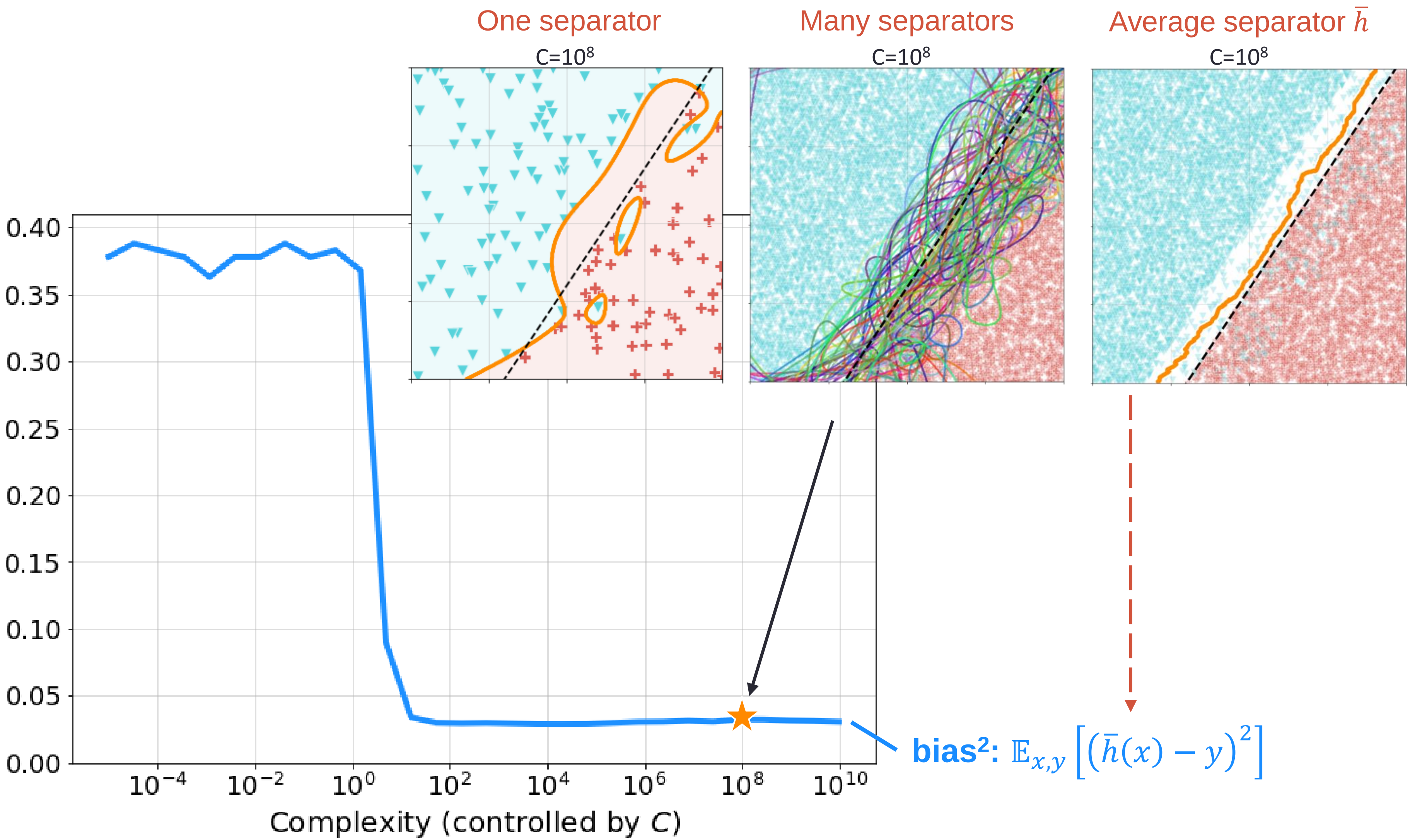


Formalizing the bias

- **Bias:**
  - Quantifies how well our hypothesis <u>class</u> fits the data (on average)
  - Formally defined as: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$
  - Does not depend on sampled data (but does depend on data size)

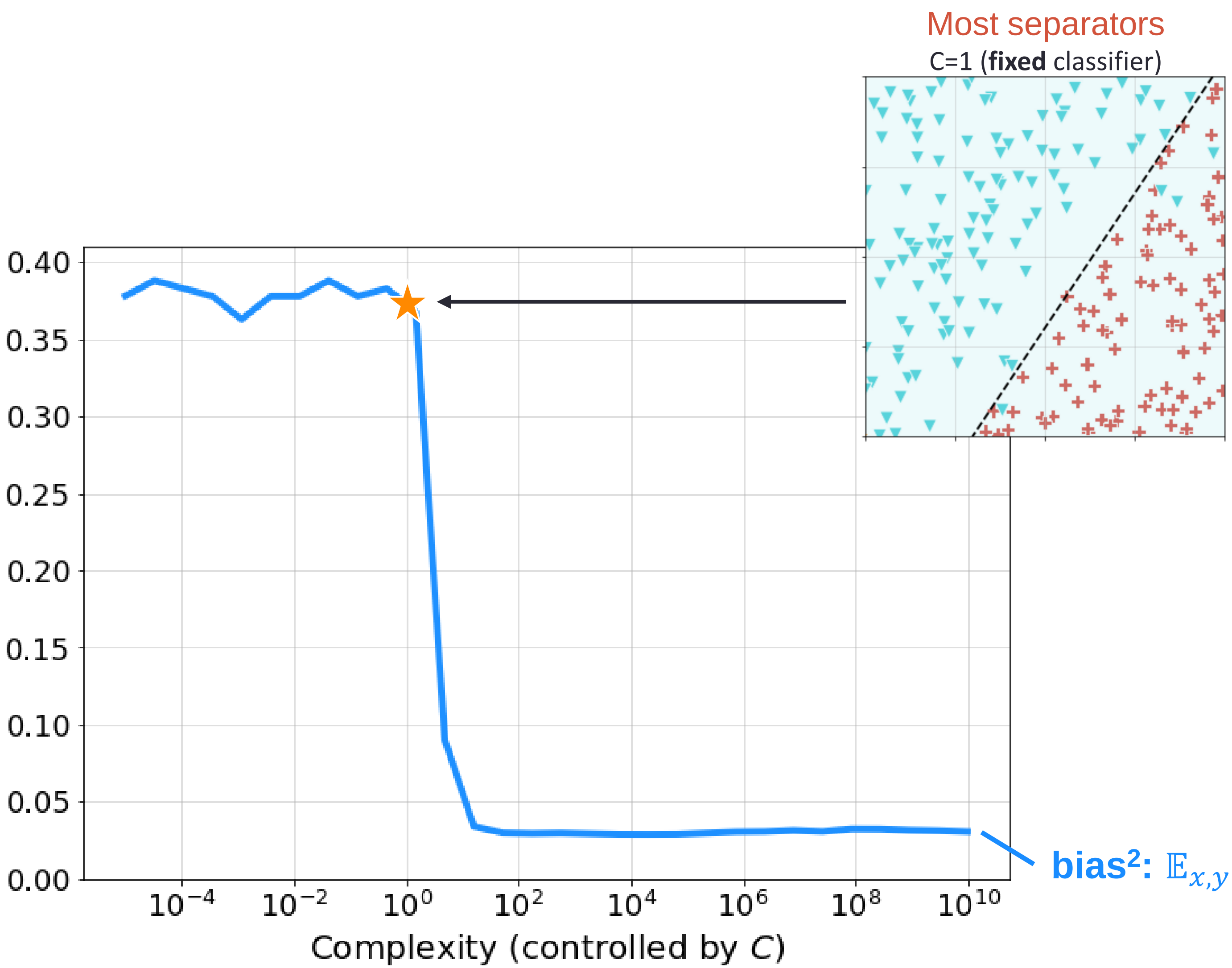


More complex models $\Longrightarrow$ less bias



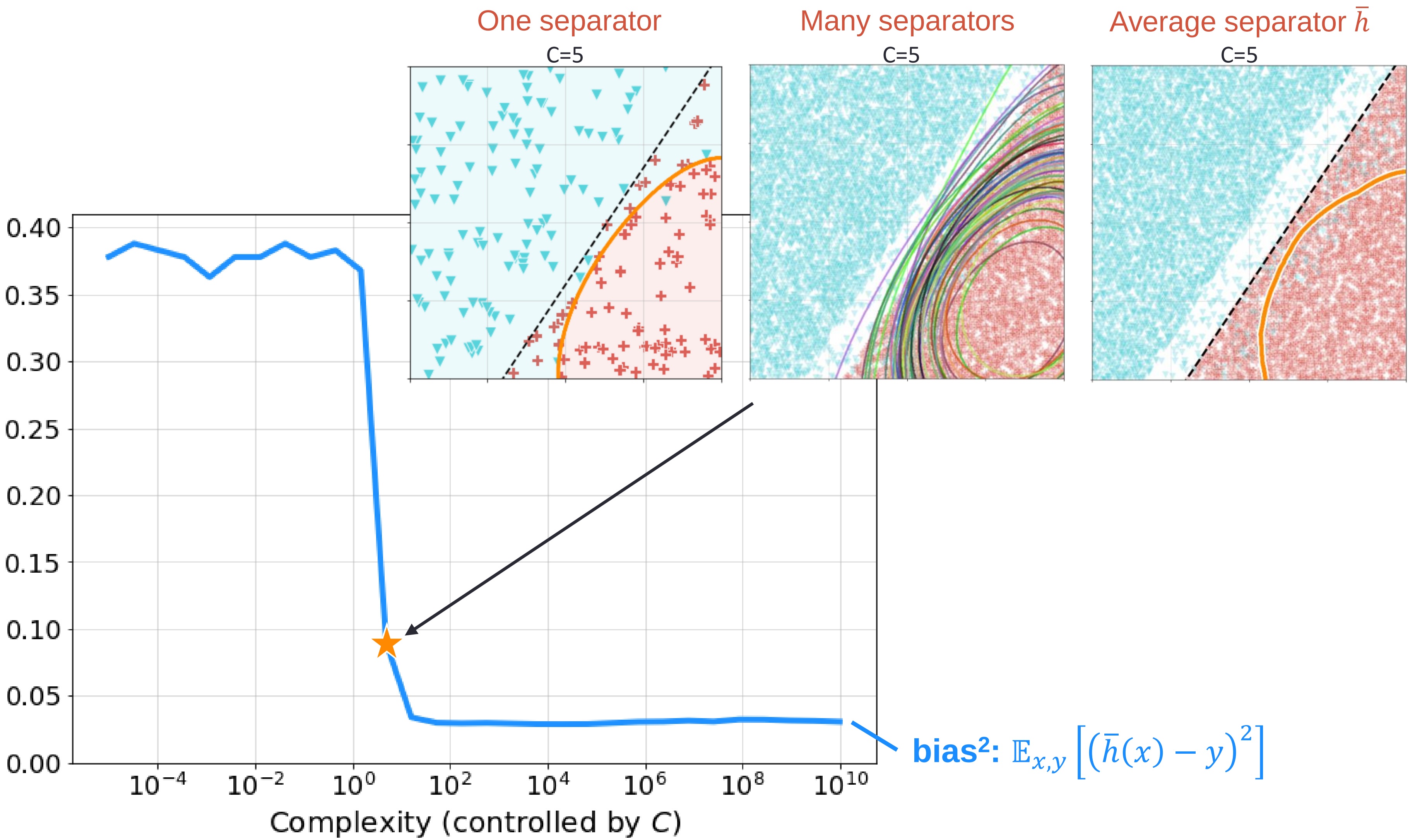
bias²: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x)-y\right)^2\right]$

More complex models $\Longrightarrow$ less bias



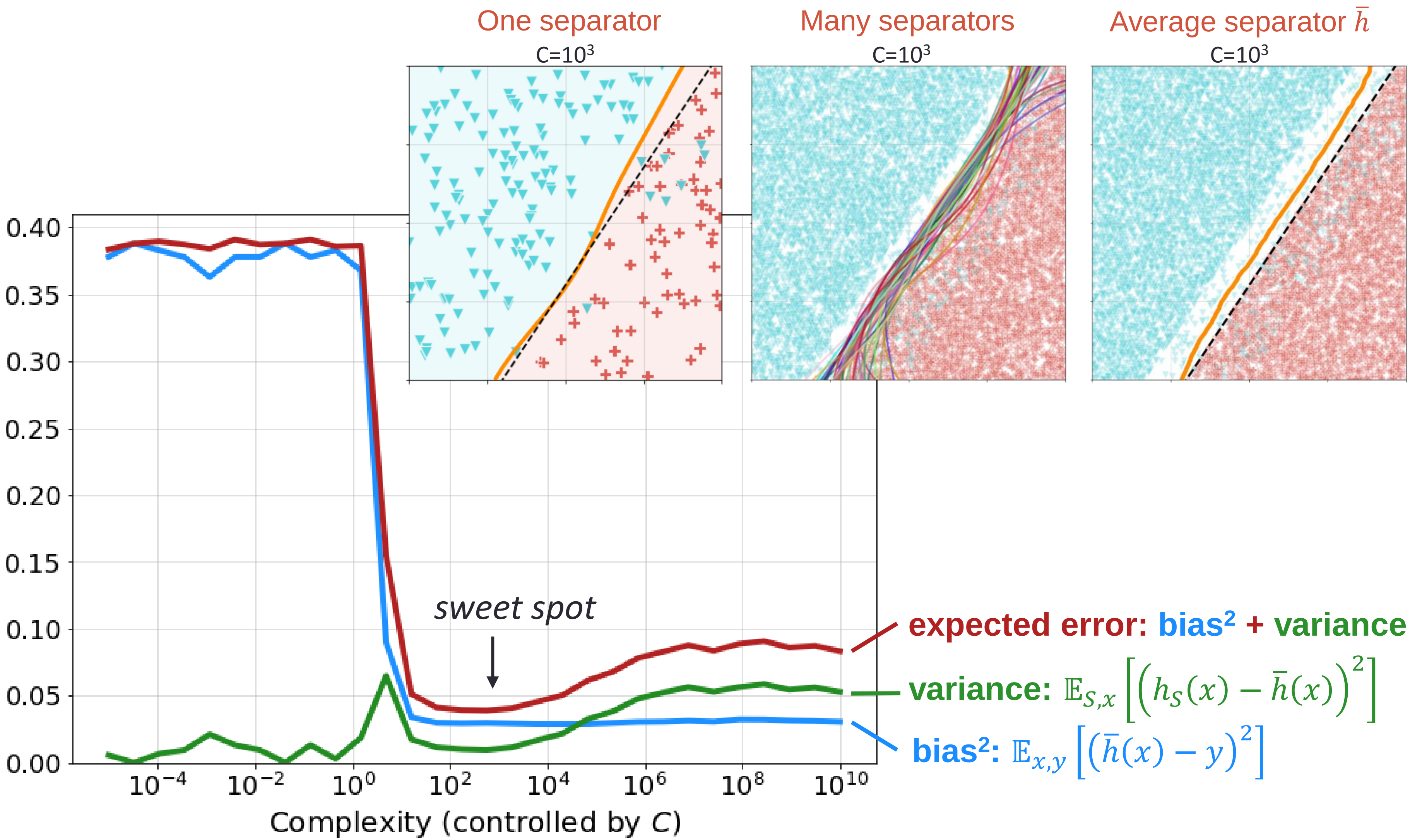
bias²: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$

More complex models ⟹ less bias



bias²: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$

The sweet spot of the expected error

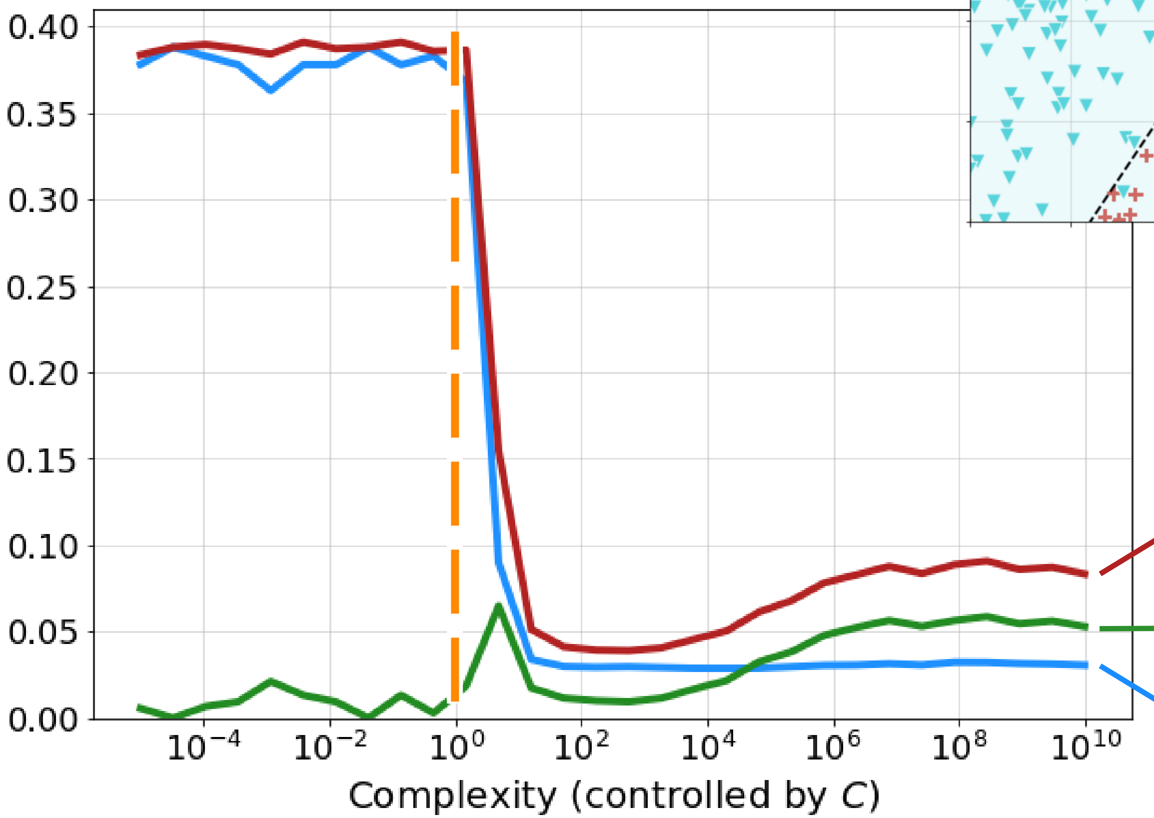


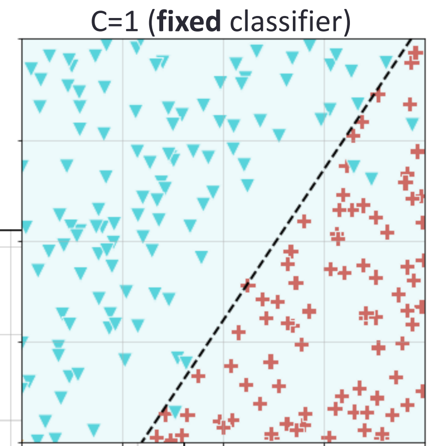
expected error: bias^2 + variance

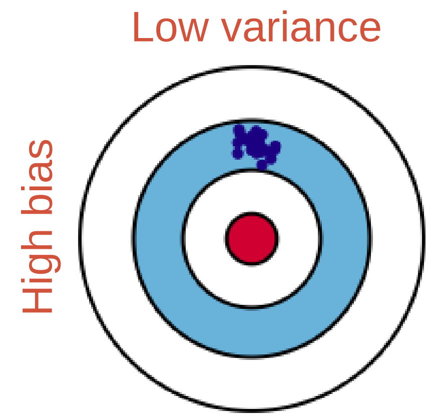
variance: $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$

bias^2: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$

Cases we saw





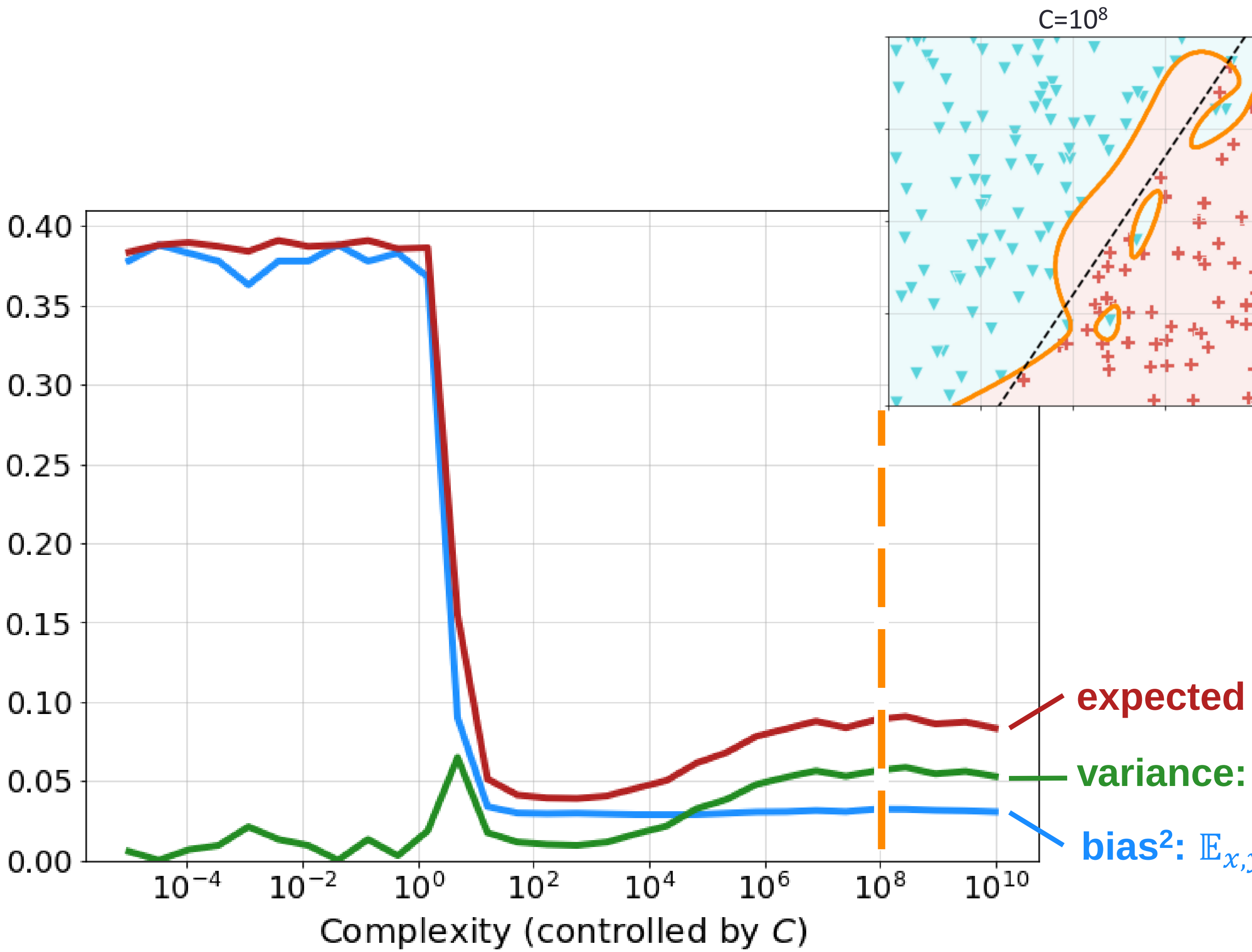


expected error: bias² + variance

variance: $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$

bias²: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$

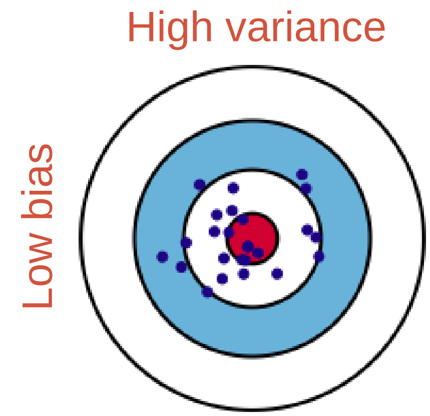
Cases we saw



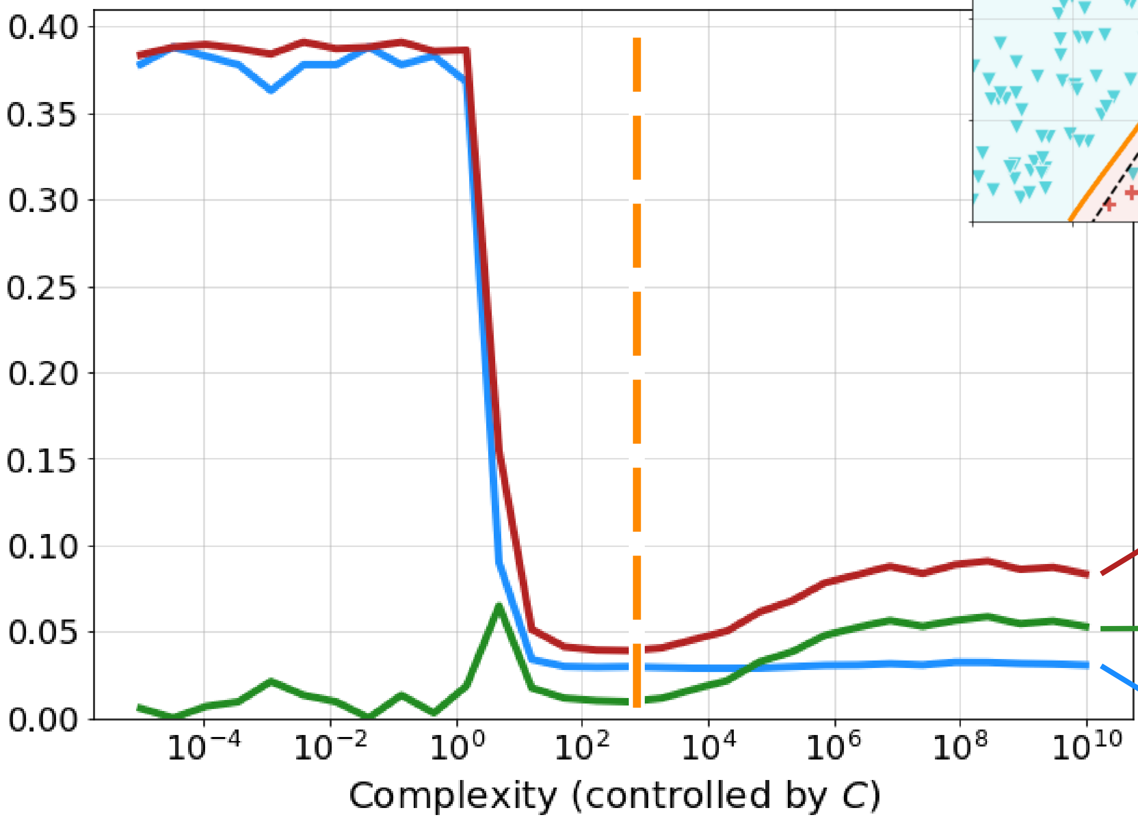
expected error: bias² + variance

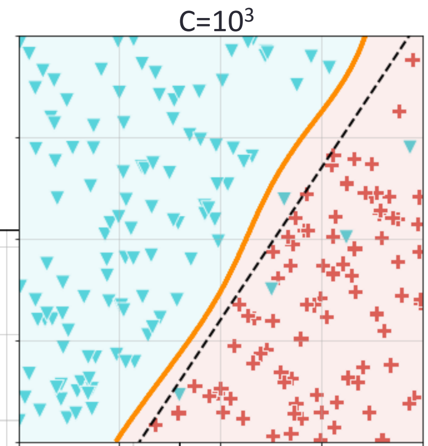
variance: $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$

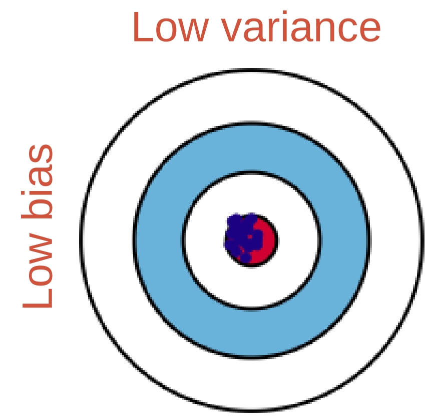
bias²: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$



Cases we saw





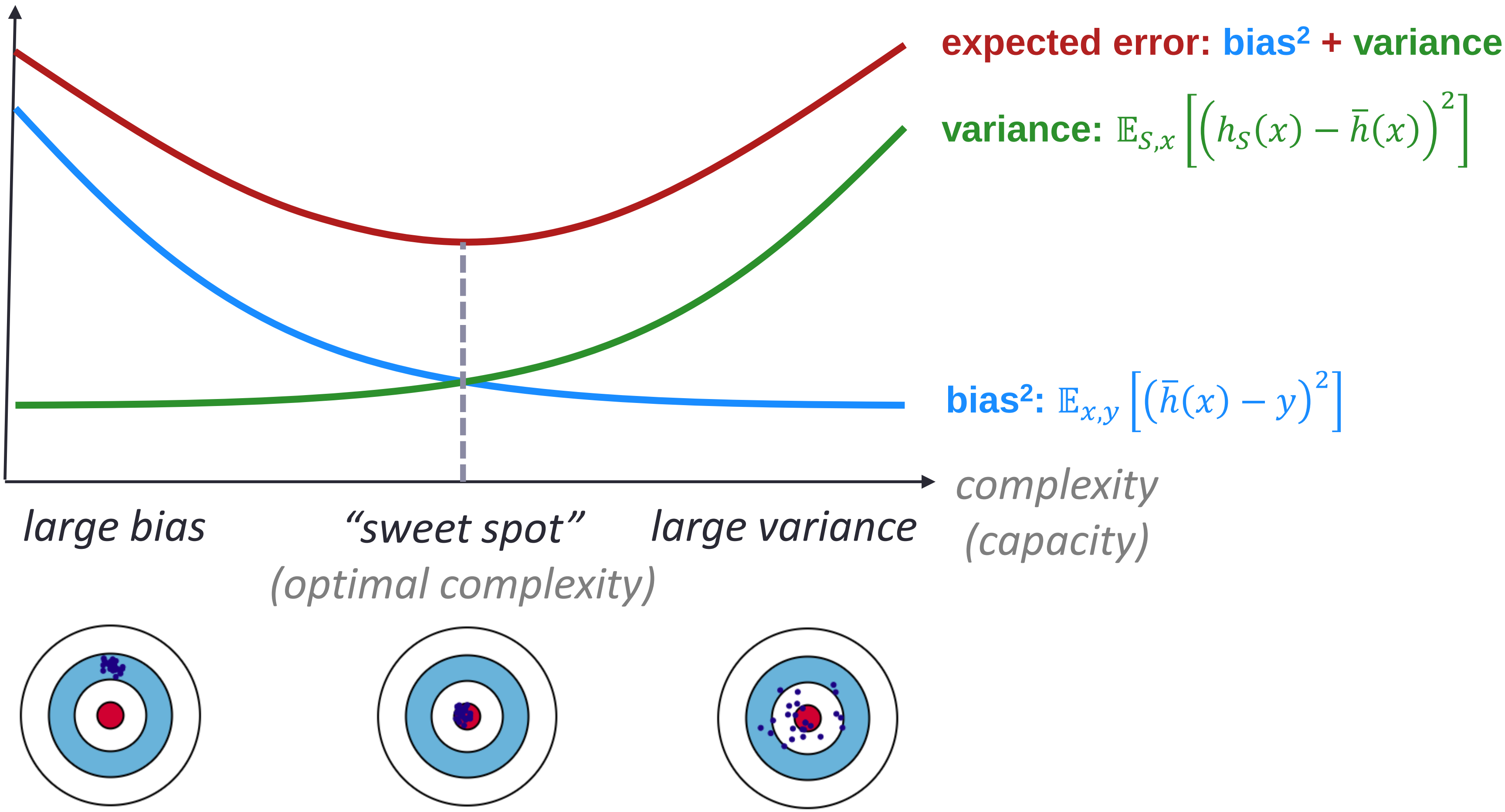


expected error: bias^2 + variance

variance: $\mathbb{E}_{S,x}\left[\left(h_S(x) - \bar{h}(x)\right)^2\right]$

bias^2: $\mathbb{E}_{x,y}\left[\left(\bar{h}(x) - y\right)^2\right]$

Bias-variance tradeoff



MODEL SELECTION

Bias-variance error decomposition

- Three interpretable sources of error:

$$\underbrace{\mathbb{E}_{S\sim D^m}\left[L_D^{sqr}(h_S)\right]}_{\textbf{\textit{expected error}}} = \underbrace{\mathbb{E}_x\left[\left(\bar{h}(x)-\bar{y}(x)\right)^2\right]}_{\textbf{\textit{bias}}^2} + \underbrace{\mathbb{E}_{S,x}\left[\left(h_S(x)-\bar{h}(x)\right)^2\right]}_{\textbf{\textit{variance}}} + \underbrace{\mathbb{E}_{x,y}\left[(\bar{y}(x)-y)^2\right]}_{\textbf{\textit{noise}}}$$

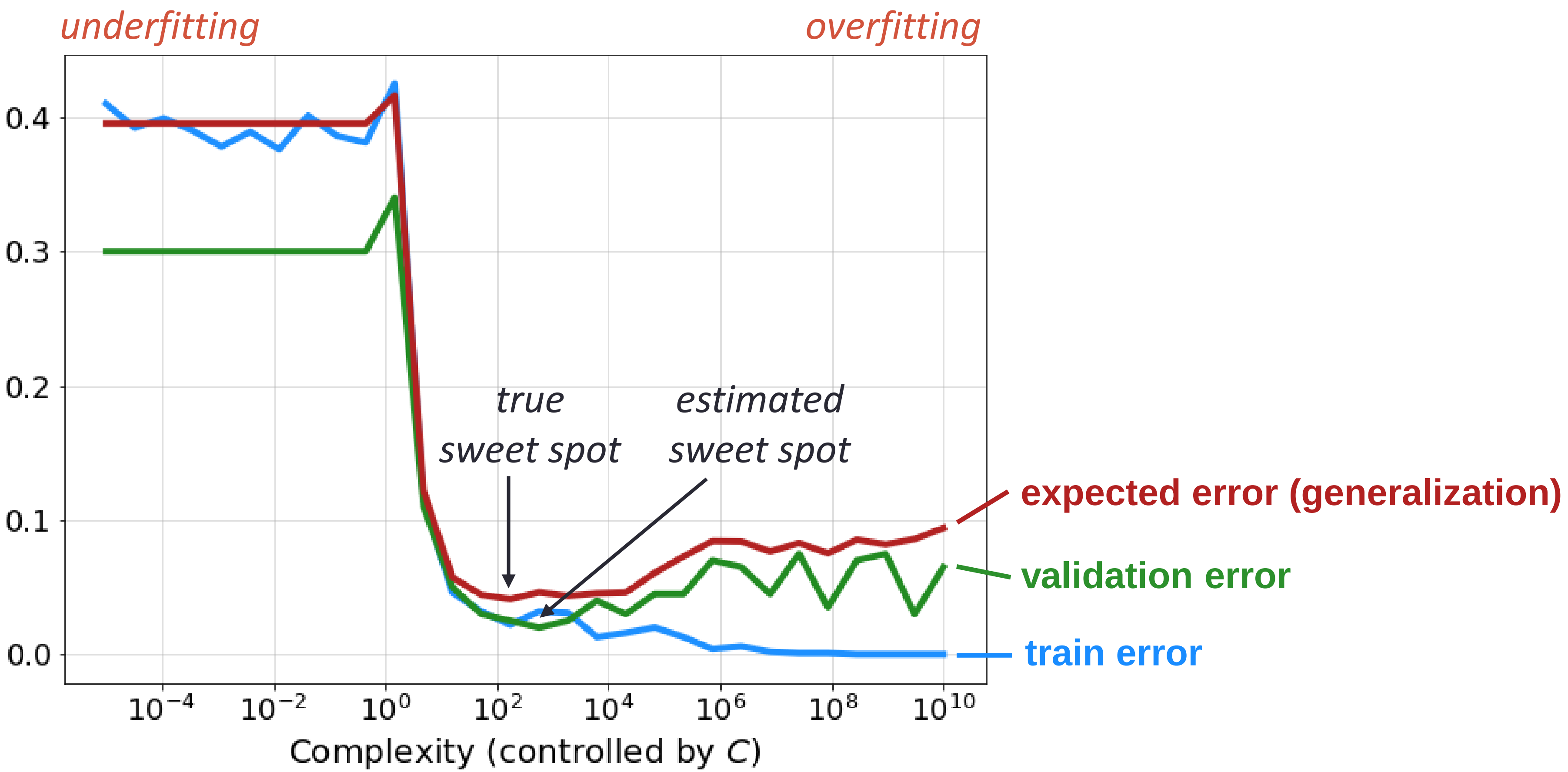
- In practice, we want to perform model selection:

  tune the model complexity to achieve a low expected error

- However, we cannot compute the actual expected error!

Validation curve

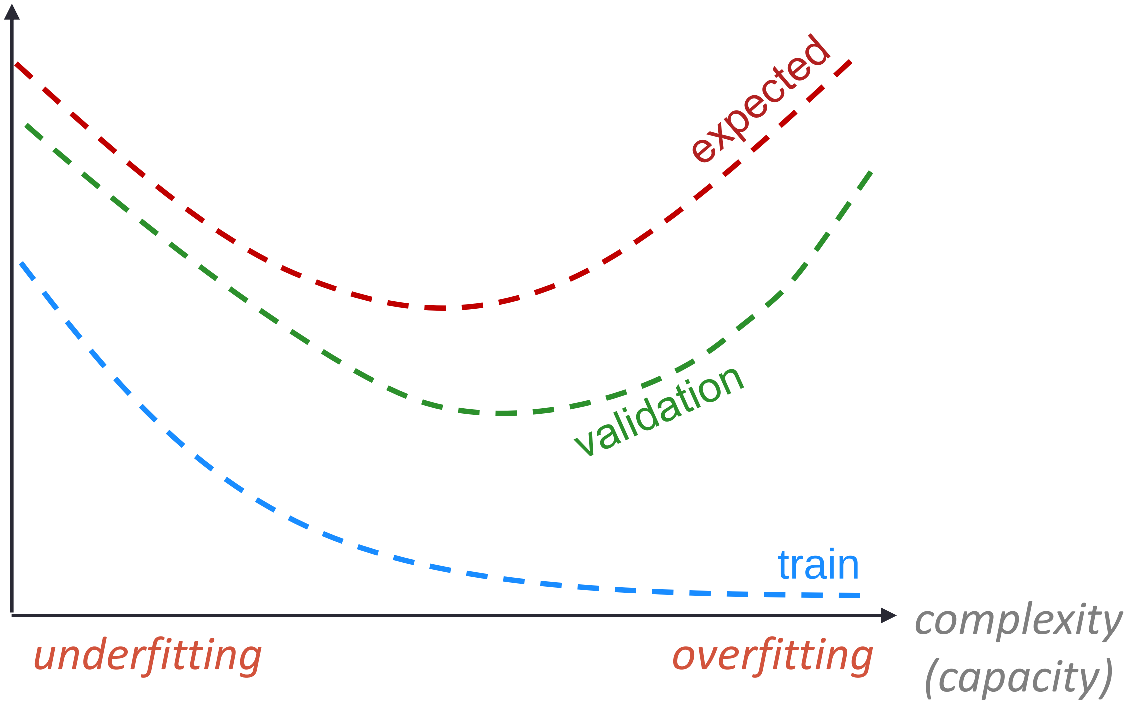
- Here, we train on 100 examples and put 20 examples aside for validation
- The validation error is an estimator for the generalization error



Error decomposition using validation

- Given a training set S, a validation set V, and a hypothesis h_S, the generalization error can be decomposed into:

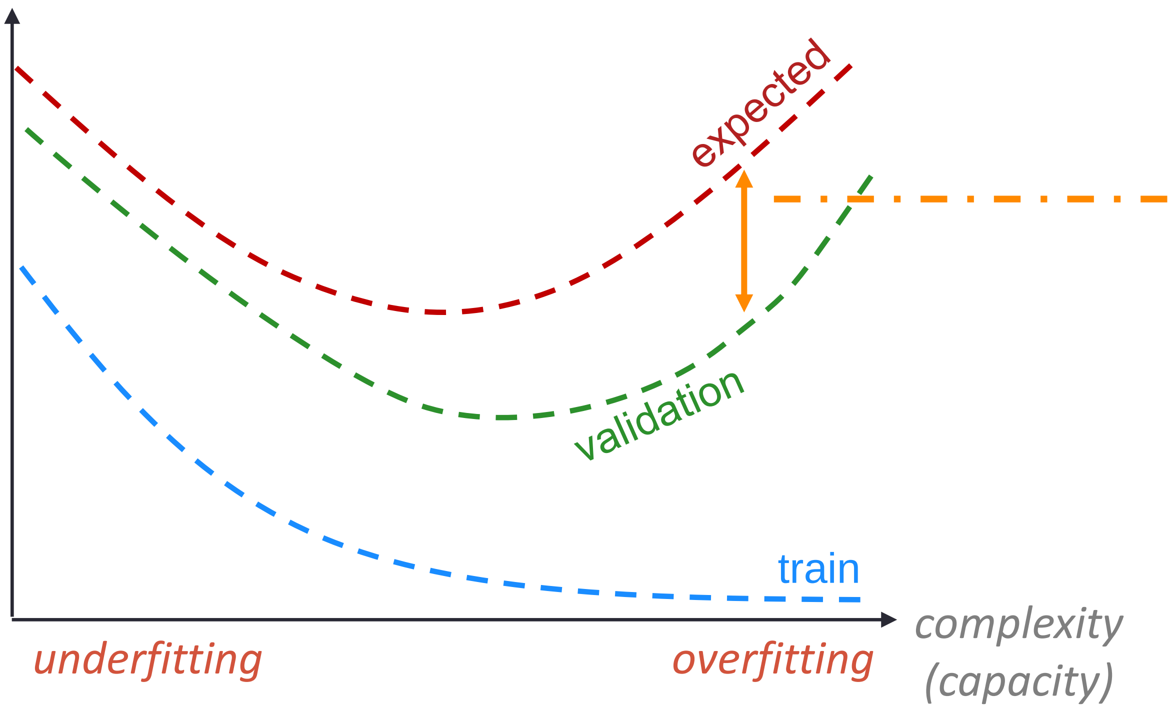
$$L_D(h_S) = \big(L_D(h_S) - L_V(h_S)\big) + \big(L_V(h_S) - L_S(h_S)\big) + L_S(h_S)$$



Error decomposition using validation

- Given a training set S, a validation set V, and a hypothesis h_S, the generalization error can be decomposed into:

$$L_D(h_S) = \underline{\left(L_D(h_S) - L_V(h_S)\right)} + \left(L_V(h_S) - L_S(h_S)\right) + L_S(h_S)$$



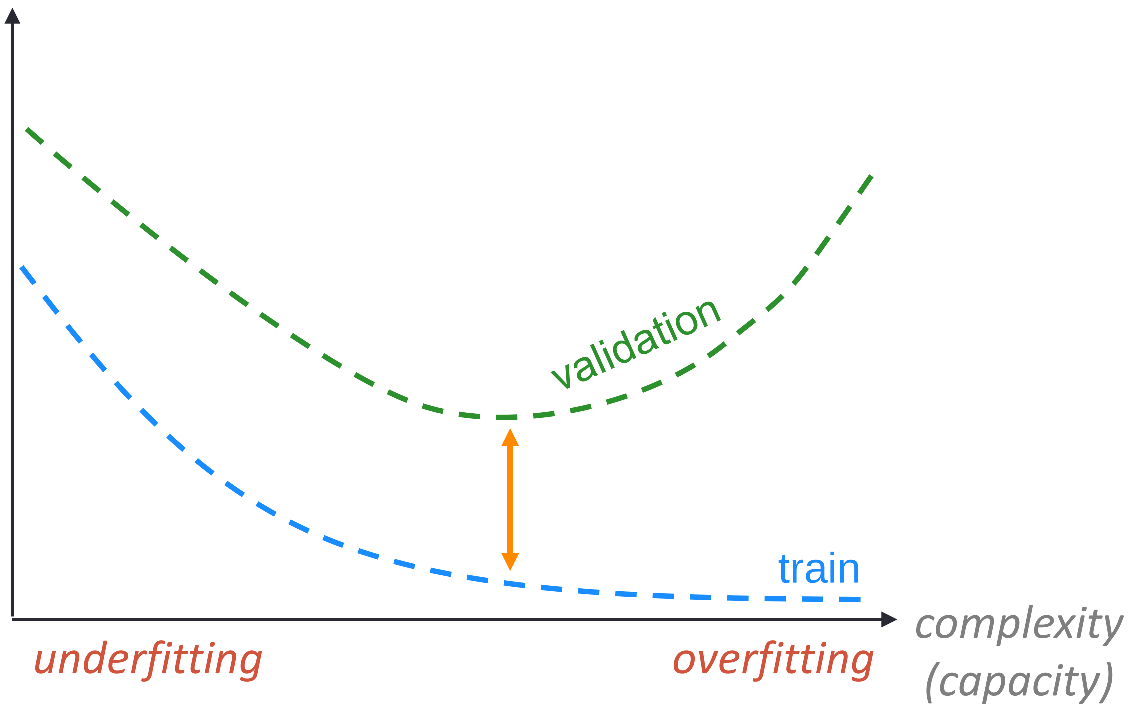
$|L_D(h_S) - L_V(h_S)|$
(bounded in the lecture)

Remember the expected error is unknown!

Error decomposition using validation

- Given a training set S, a validation set V, and a hypothesis h_S, the generalization error can be decomposed into:

$$L_D(h_S) = \big(L_D(h_S) - L_V(h_S)\big) + \underline{\big(L_V(h_S) - L_S(h_S)\big)} + L_S(h_S)$$



If this term is large,
h_S probably overfits.

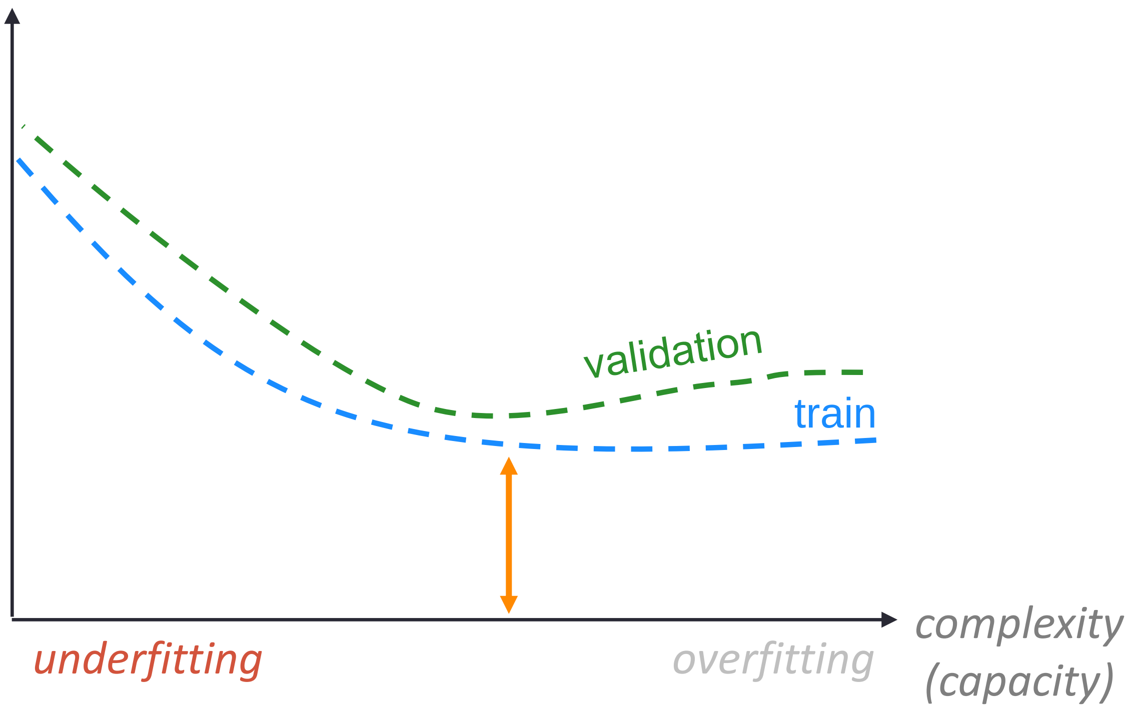
Possible solutions:

- Get more samples
- Feature selection
- Lower the capacity
- Change regularization type

Error decomposition using validation

- Given a training set S, a validation set V, and a hypothesis h_S, the generalization error can be decomposed into:

$$L_D(h_S) = \big(L_D(h_S) - L_V(h_S)\big) + \big(L_V(h_S) - L_S(h_S)\big) + \underline{L_S(h_S)}$$



If this term is large,
h_S probably underfits.

Possible solutions:

- Increase the complexity
- Improve tuning
- Change feature mapping
- Change hypothesis class

Summary

- The model complexity creates a tradeoff between bias and variance
- Model selection
  - Validation curves help tune hyperparameters (and model complexity)
  - Error could also be decomposed using validation
  - Further reading: Chapter 11 in Understanding ML: From Theory to Algorithms