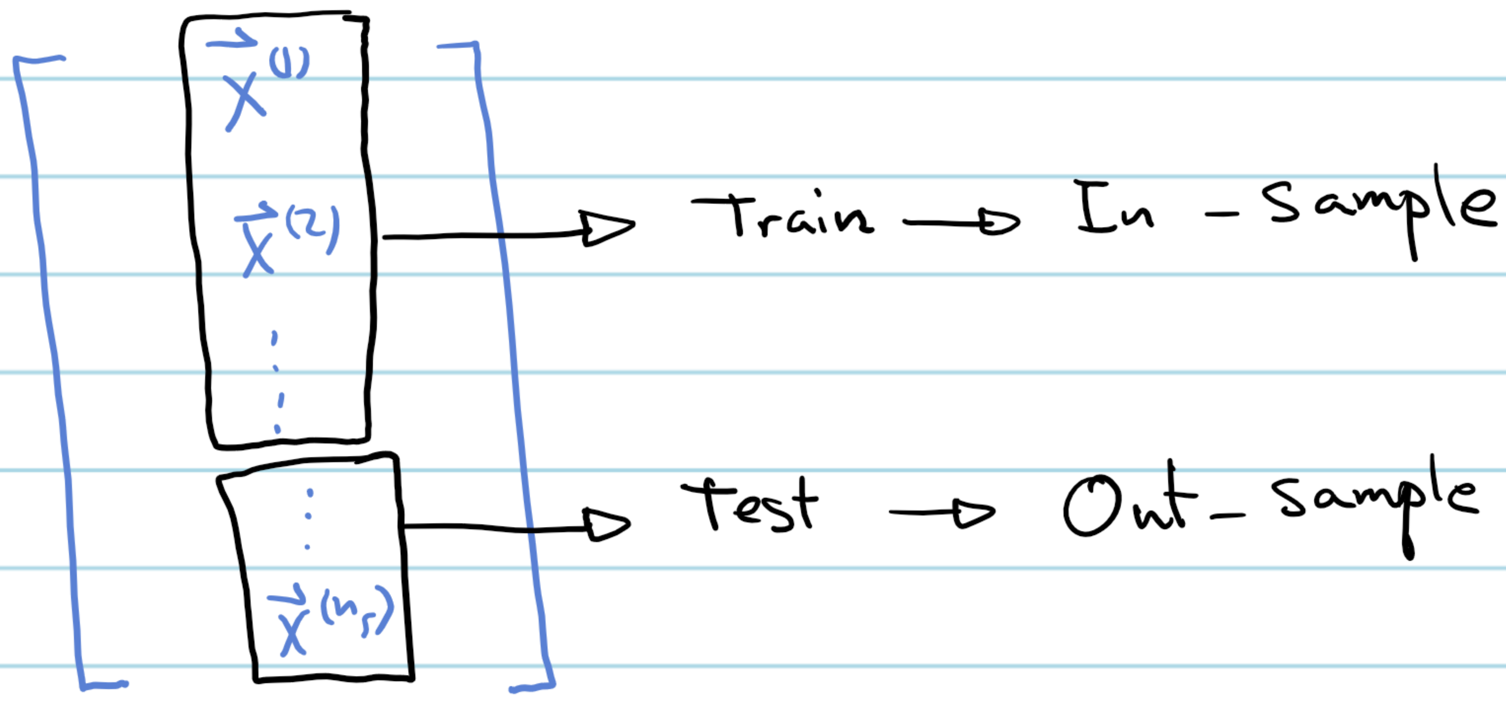
In & Out – sample performance

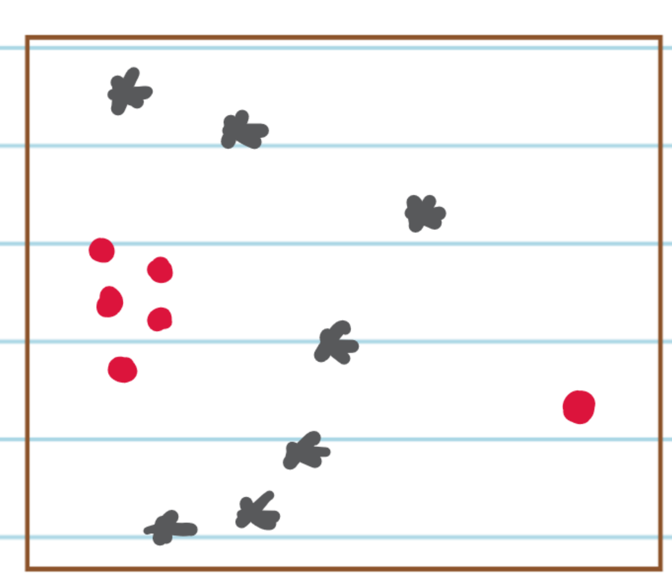


Ideally, we want models to have high performance on both train & test data.

A critical assumption here is that both test & train set come from the same distribution.

→ So we need to shuffle the data before splitting the data into train & test.

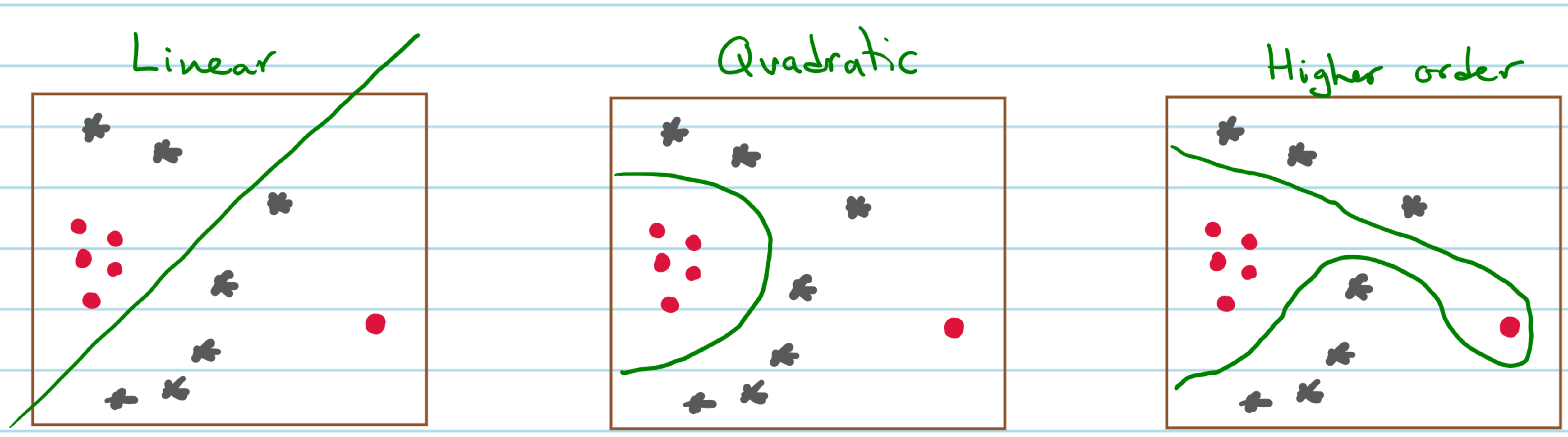
Overfitting & Underfitting



Consider the data in this picture.

There are two classes (red & gray).

Consider the following classifications:



Under_fit
High Bias

Over_fit
High Variance.

Bias vs Variance

* Some terminology & notation (from statistical learning)

Hypothesis set $\mathcal{H}$: All the possible functions that we may choose our model from.

Example: Decision trees based on diff. questions

Polynomial models. $f(x) = \text{poly}(x)$

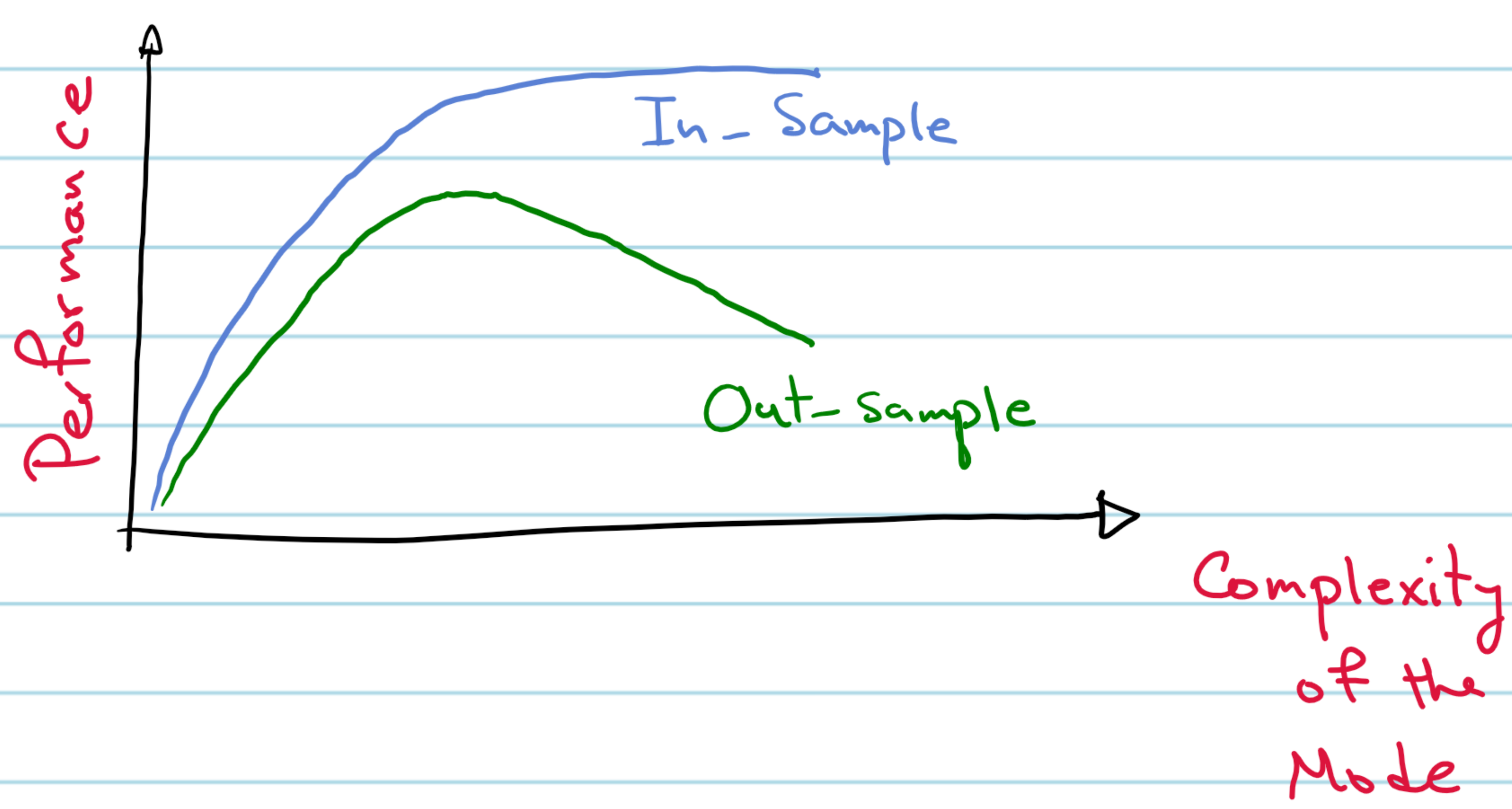
Data: A set of points $\{x_i, f(x_i)\}$ for $i = 1 \ldots N$ (sampled from the same distribution i.e. f).
function

In- & Out- Sample error: Based on some metric, the error rate on the training data (used to pick the model from hypothesis set) & test data (new unseen data).

How do you guess these models do in each category?

	Linear	Quadratic	Higher orders
In-Sample accuracy	Low	High	Very high
Out-Sample accuracy	Low (close to in-sample)	High (close to in-sample)	Low

Typically



This give a notion of where to stop.

This plot is known as the "validation curve".

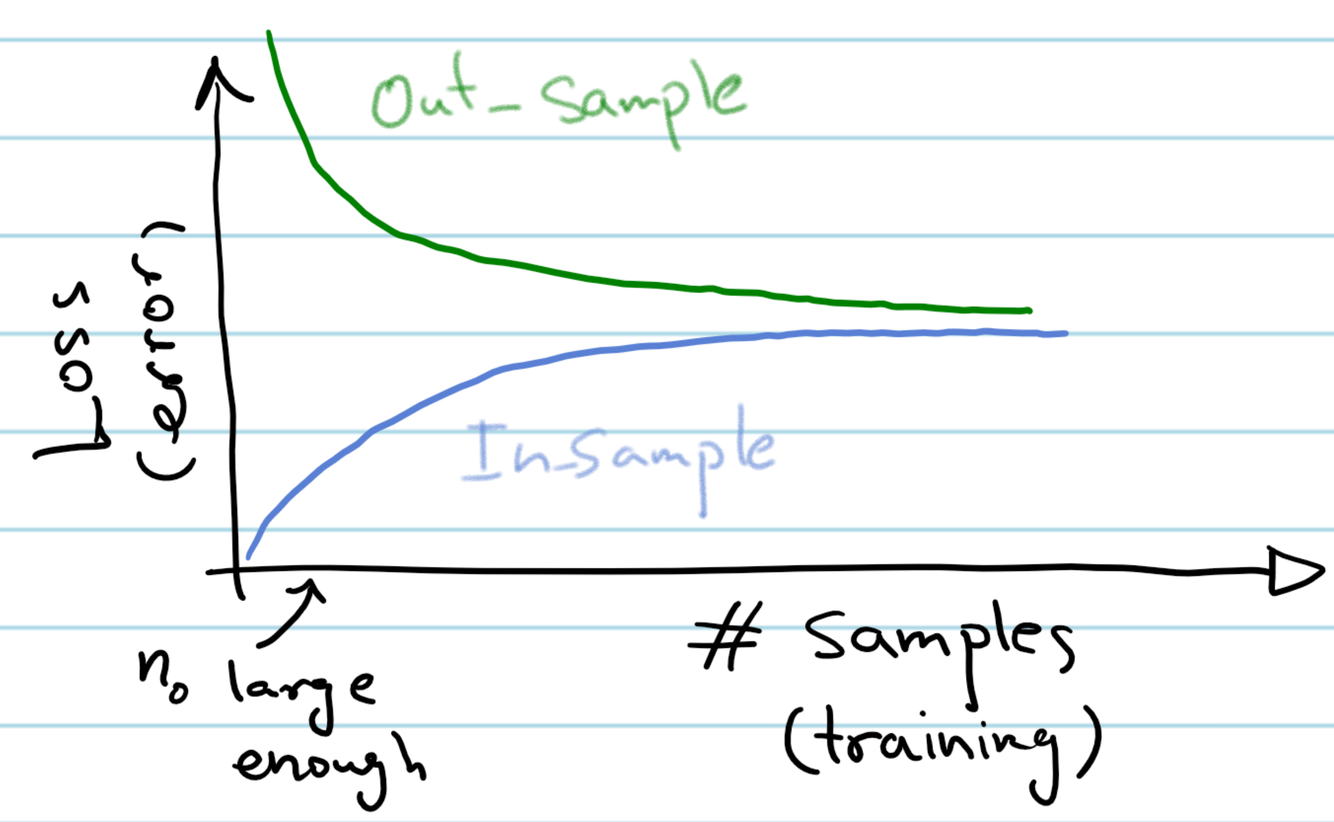
Learning Curve

One of the best ways to avoid overfitting is to use more data.

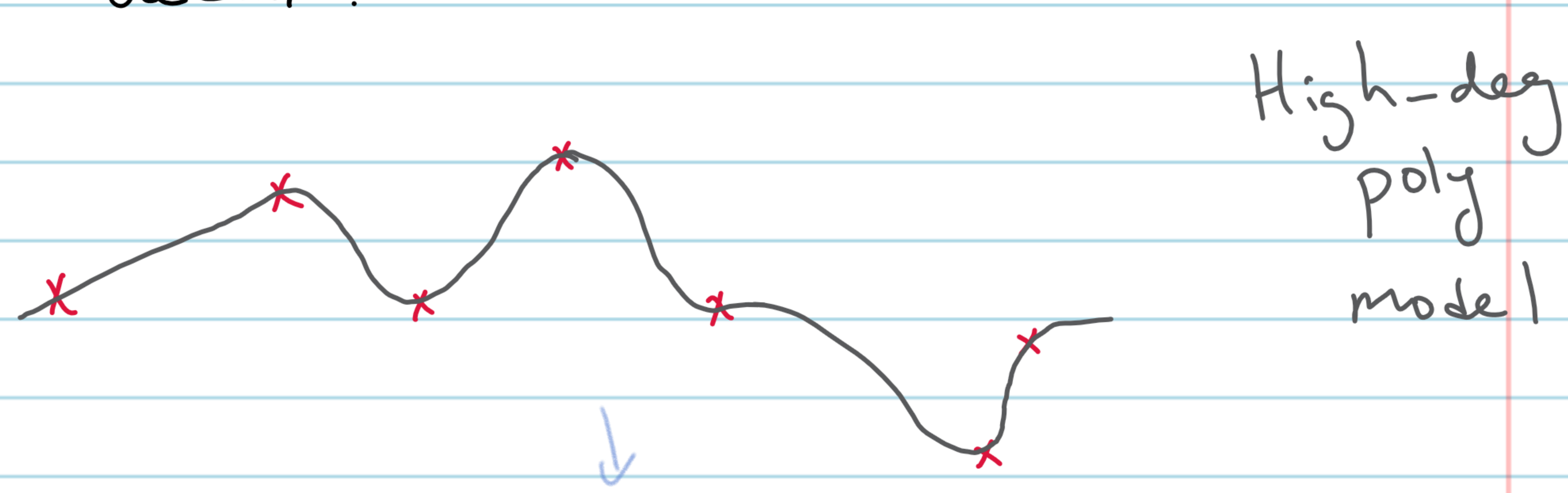
But more data does not always help with improving the performance.

Here we introduce some tools to help us get a sense of when we need more data and also how good our model is.

Learning Curve

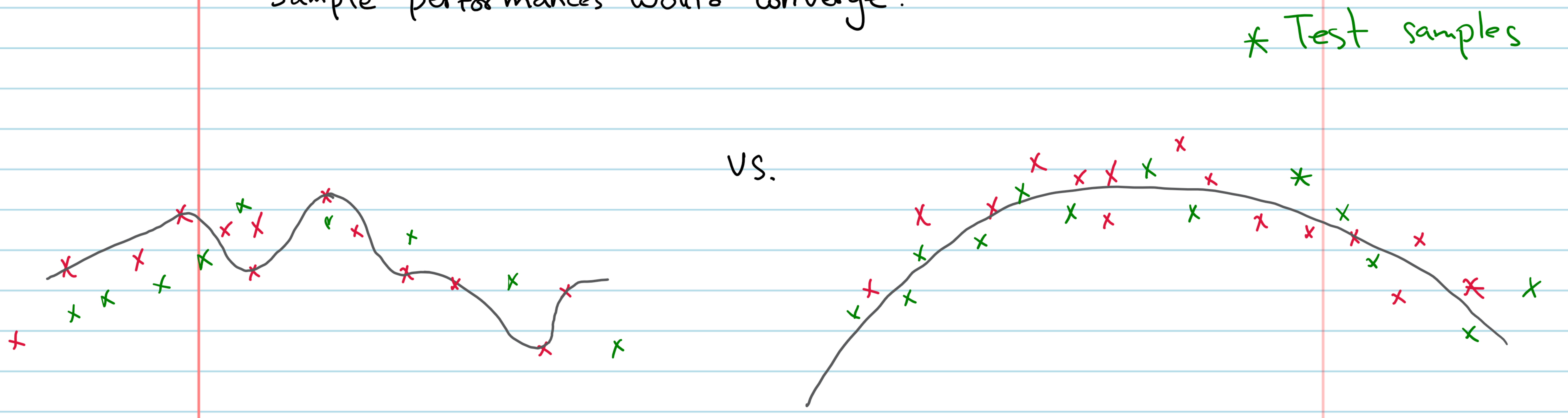


With more training samples, the performance over the training set would decline:



With more points it is not possible to capture all the details.

On the other hand, with more data, the training process would start to focus on the major trends rather than small fluctuations and the in- & out-sample performances would converge.



- This brings us to the concepts of Bias & Variance.