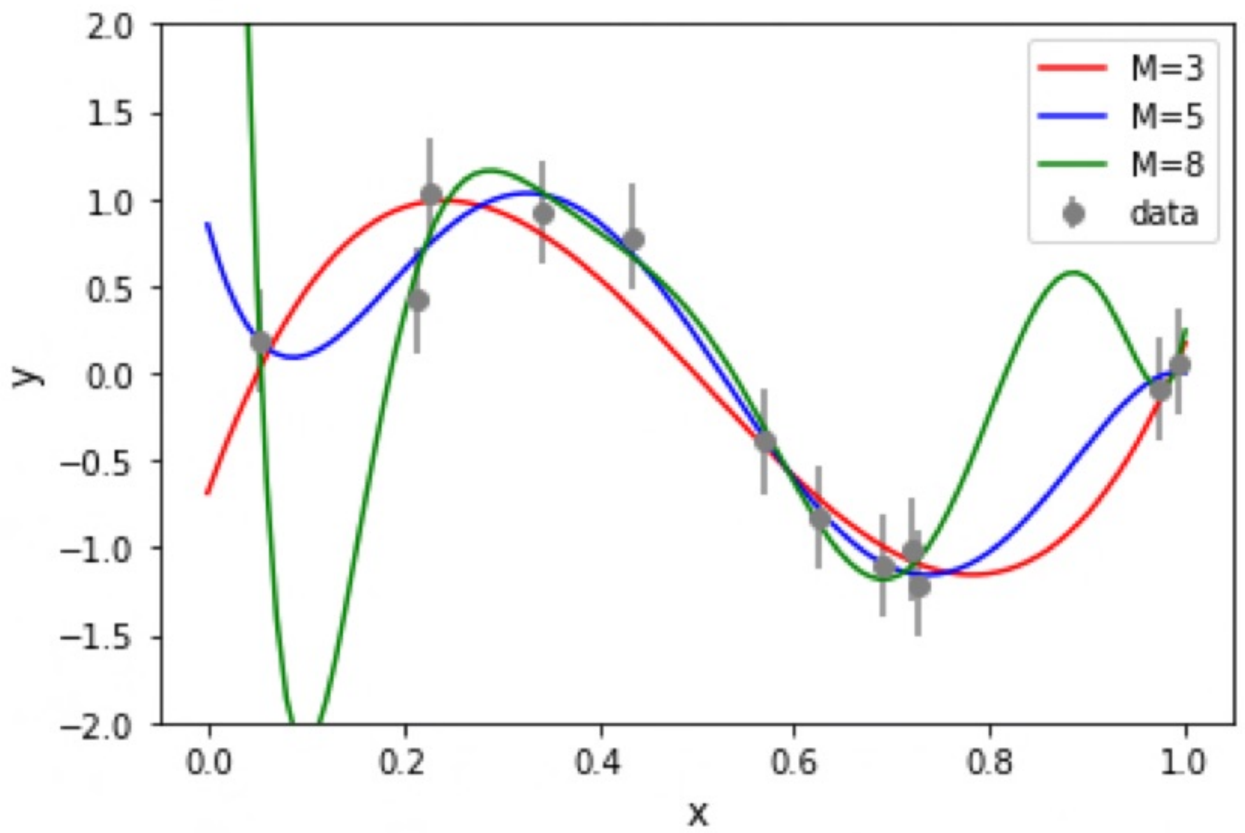
• Model selection (both freq and Bayesian)

By looking at this plot what would be the best model?

3rd-order polynomial

5th-order one

or 8th-order one?



Note: model M=8 fits the data better than the rest!

Intuitively model M=8 is not the best though. This is because it seems like if M=8 doesn't look like [the unknown function which generated this data]

↳ this is ultimately what we are after in statistical inference!

So if a new point appears after the fitting is done, it will probably be poorly predicted by M=8

M=8 is said to overfit the data

Somehow the whole problem of model selection is to find the model which describes the data well (i.e. not doing underfitting either) while not doing overfitting

* One of the most common ways to do model comparison in physics is by computing the

$$\left[\chi^2 / \text{d.o.f.} \right]$$

The justification for that is based on what we saw previously about using χ^2 as a goodness-of-fit:

Assume that observations $\{y_i\}$ follow a Gaussian distribution with mean

$$\mu_i \equiv \mu(x_i; \vec{\theta})$$

$\vec{\theta}$ → a set of m parameters

and known std's σ_i } e.g. the errors of the measurements

then the quantity $\chi^2 = \sum_{i=1}^{N} \left(\frac{y_i - \mu_i}{\sigma_i} \right)^2$

follows a χ^2-distrib with $N-m$ d.o.f.

So χ^2 is a measure of how likely A is that data comes from the model whose predictions are the μ_i, in units of the stds.

On the other hand, noting that the mean of χ^2 distrib. is equal to the DOF, then the quantity

rule of thumb { χ^2/DOF :

- if $\gg 1$: bad model, underfitting
- if ~ 1 : fit is reasonable, within 1 std
- if $\ll 1$: overfitting, or the errors have been overestimated

This rule-of-thumb has several disadvantages:

1) if two models give

$$\left(\chi^2/\text{DOF}\right)_1 \approx \left(\chi^2/\text{DOF}\right)_2 \approx 1$$

Then one cannot really favour one model against the other

2) Suppose model #1 gives $\chi^2 = 15$ and DOF = 10,
whereas " #2 " $\chi^2 = 150$ and DOF = 100;

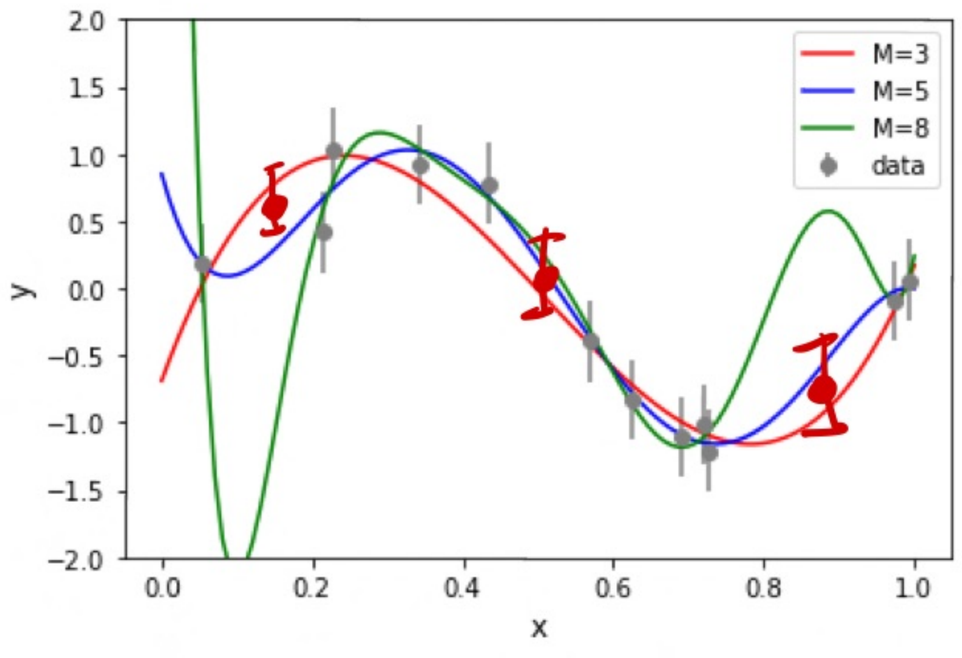
↳ both have the same χ^2/DOF, however

$$(\text{p-value})_1 = 0.13 \,, \quad (\text{p-value})_2 = 9\times10^{-4}$$

3) If data is not normally distributed, then the conclusions may be incorrect.

* Another frequentist method, probably the most widely used by the ML community, is to fit the model with part of the data ("training data" in the ML jargon) and then evaluate how good the fitted model performs in the rest of the data ("test data")

clearly the M=8 performs the worst regarding the test data points ⇐



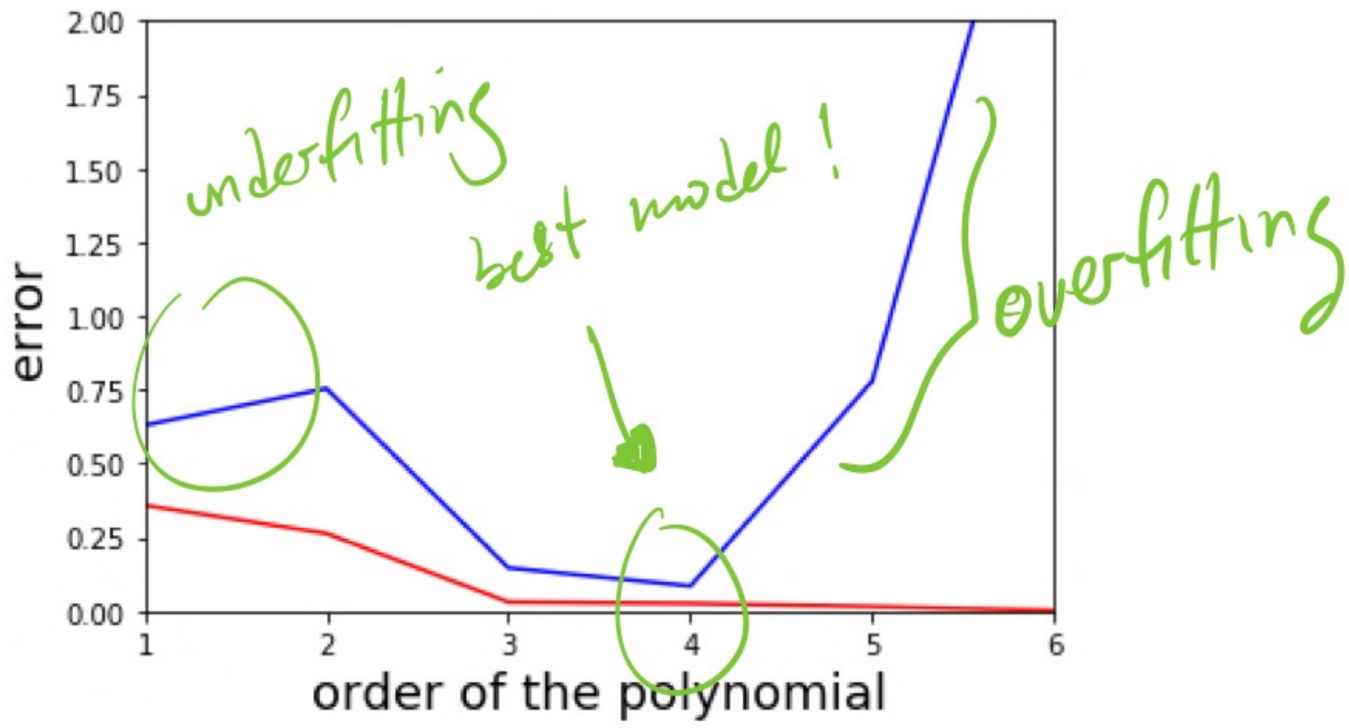
This is also an heuristic method, but it is much more powerful than the χ^2/DOF in many ways; e.g.:

- By computing the discrepancy between data and predictions, using the likelihood

$$p(y_i|\vec{\theta})$$

one can choose the model which gives the lowest discrepancy (the largest likelihood) on the test data

- The method is not restricted to normally distributed data!



- Drawback: maybe not convenient for small datasets: one ideally would like to use enough "training" data for the fit to be good, but it may be insufficient in this case

- Note that when a model underfits, the bias (diff. between prediction and data) is large, while

the variance is small (predictions don't change much from point to point)

On the contrary, if a model overfits, the bias is small but the variance is high.

⇒ So a good model has a "Bias-Variance" trade-off.

Actually, it can be shown that the MSE

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

is a linear combination of bias and variance:

Let $y(x) = f(x) + \varepsilon$ { $f(x)$ is the true underlying function we are trying to model with $\hat{f}(x)$

$$\varepsilon \sim N(0, \sigma^2)$$

$$Var[\hat{f}] = \mathbb{E}[\hat{f}^2] - (\mathbb{E}[\hat{f}])^2$$

$$Bias[\hat{f}] = \mathbb{E}[\hat{f}] - f$$

} $\mathbb{E}$ is taken wrt different datasets one could have access to

Actually:

$$MSE = \mathbb{E}[(y - \hat{f}(x))^2]$$

↳ new point, unseen before ("test" point)

$$= \mathbb{E}[(f + \varepsilon - \hat{f})^2]$$

$$= \mathbb{E}[(f + \varepsilon - \hat{f} + \mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}])^2]$$

$$= \mathbb{E}\Big[\big(f - \mathbb{E}[\hat{f}]\big)^2 + \mathbb{E}[\varepsilon]^2 + \big(\mathbb{E}[\hat{f}] - \hat{f}\big)^2 + 2\big(f - \mathbb{E}[\hat{f}]\big)\varepsilon + 2\big(\mathbb{E}[\hat{f}] - \hat{f}\big)\varepsilon + 2\big(f - \mathbb{E}[\hat{f}]\big)\big(\mathbb{E}[\hat{f}] - \hat{f}\big)\Big]$$

$$= \big(f - \mathbb{E}[\hat{f}]\big)^2 + \mathbb{E}[\varepsilon]^2 + \mathbb{E}\big[\big(\mathbb{E}[\hat{f}] - \hat{f}\big)^2\big] + 2\Big\{\big(f - \mathbb{E}[\hat{f}]\big) + \mathbb{E}\big[\mathbb{E}[\hat{f}] - \hat{f}\big]\Big\}\mathbb{E}[\varepsilon] + 2\big(f - \mathbb{E}[\hat{f}]\big)\,\mathbb{E}\big[\mathbb{E}[\hat{f}] - \hat{f}\big]$$

$\mathbb{E}[\varepsilon] \to 0$

$\mathbb{E}[\hat{f}] - \mathbb{E}[\hat{f}] = 0$

$$= \underbrace{\big(f - \mathbb{E}[\hat{f}]\big)^2}_{\text{Bias}^2} + \underbrace{\mathbb{E}[\varepsilon]^2}_{\text{Var}[\varepsilon] + (\mathbb{E}[\varepsilon])^2,\ 2\text{Var}[\varepsilon] = \sigma^2} + \underbrace{\mathbb{E}\big[\big(\mathbb{E}[\hat{f}] - \hat{f}\big)^2\big]}_{\text{Var}}$$

$$= \text{Bias}^2[\hat{f}] + \text{Var}[\hat{f}] + \sigma^2$$

* Bayesian Model Comparison

Expanded version of the Bayes theorem:

$$p(\vec{\theta} \mid \text{model}, \text{data}) = \frac{p(\text{data} \mid \vec{\theta}, \text{model})\, p(\vec{\theta} \mid \text{model})}{p(\text{data} \mid \text{model})}$$

"model evidence"

Suppose 2 models : m_1 and m_2

then $p(m_i \mid data) \propto \underbrace{p(data \mid m_i)}_{\text{evidence}} \overbrace{p(m_i)}^{\text{prior of the model itself (not of its params!)}}$

So if both models have equal prior,

$\Rightarrow$ the most probable model is obtained by computing the Bayes factor :

$$B = p(data \mid m_1) / p(data \mid m_2)$$

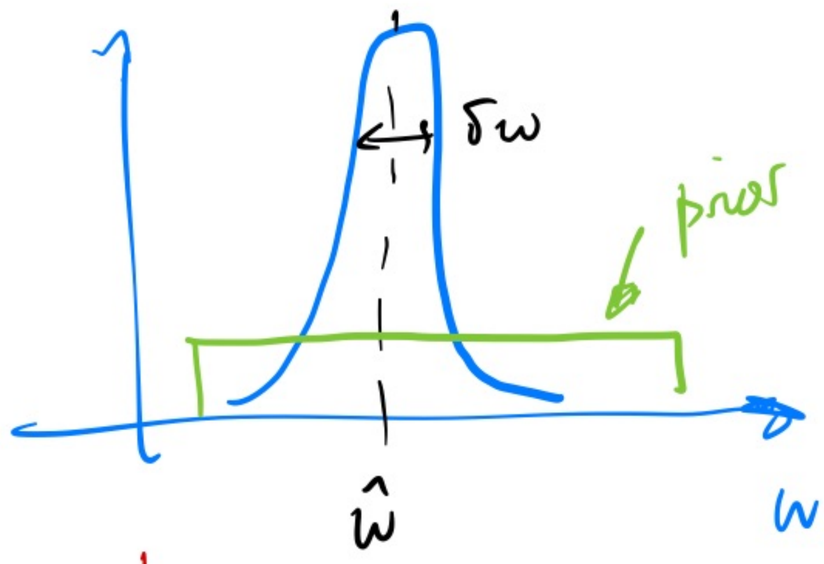
There is a criterion for B : The Jeffreys scale

B	: strenght of evidence
< 1	negative
1-3	barely worth mentioning
3-10	substantial
10-30	strong
30-100	very strong
> 100	decisive

- In order to get insight about the model evidence, let's suppose: • a model with 1 single parameter w
  - likelihood is sharply peaked around $\hat{w}$ having width δw
  - flat prior : $p(w) = \frac{1}{\Delta w_{prior}}$

then

$$p(D|m) \approx \frac{p(D|\hat{w})\,\delta w}{\Delta w_{prior}}$$



penalisation for models having large prior uncertainties

- Now assume a model with K parameters $\vec{w} = \{w_k\}$

Suppose similar ratio $\frac{\delta w_k}{\Delta w_{prior}^k} = \frac{\delta w}{\Delta w_{prior}} \quad \forall\, k$

then

$$p(D|m) \simeq p(D|\hat{\vec{w}}) \left(\frac{\delta w}{\Delta w_{prior}}\right)^K$$

penalisation for models having too many parameters, (likelihood $p(D|\hat{\vec{w}})$ improves though, so optimal model should have an intermediate complexity))

Drawback: Computing the evidence is typically intractable!
(a whole topic on approximate Bayesian inference)