

# LECTURE 3: INTRO TO DATA ANALYSIS AND MACHINE LEARNING

- **Goal of data analysis:** to determine some parameters from the data
- We want to combine new data with previous information on the parameters (prior: theoretical or empirical)
- We multiply the likelihood of parameters given the data with the prior to get the posterior
- **Goal of machine learning:** we want to predict parameters of new data given some existing labeled data (supervised learning)
- The goals of statistics and ML are often similar or related
- Methodologies and language are often very different

- **Data analysis: summarizing the posterior information:** mean or mode, and variance. Typically we are interested in more than mean and variance
- **Posterior intervals:** e.g. 95% credible interval can be constructed as central (relative to median) or highest posterior density. Typically these agree, but:

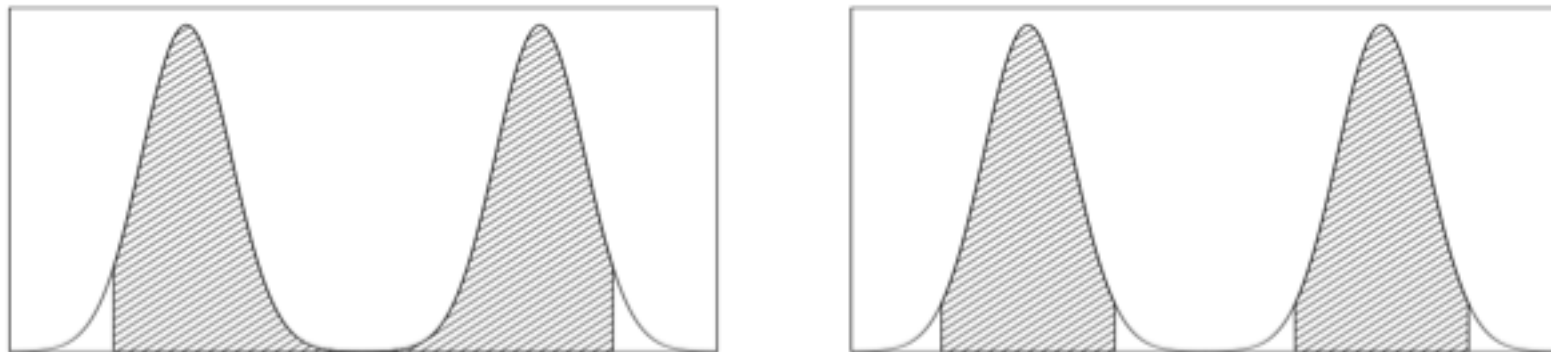
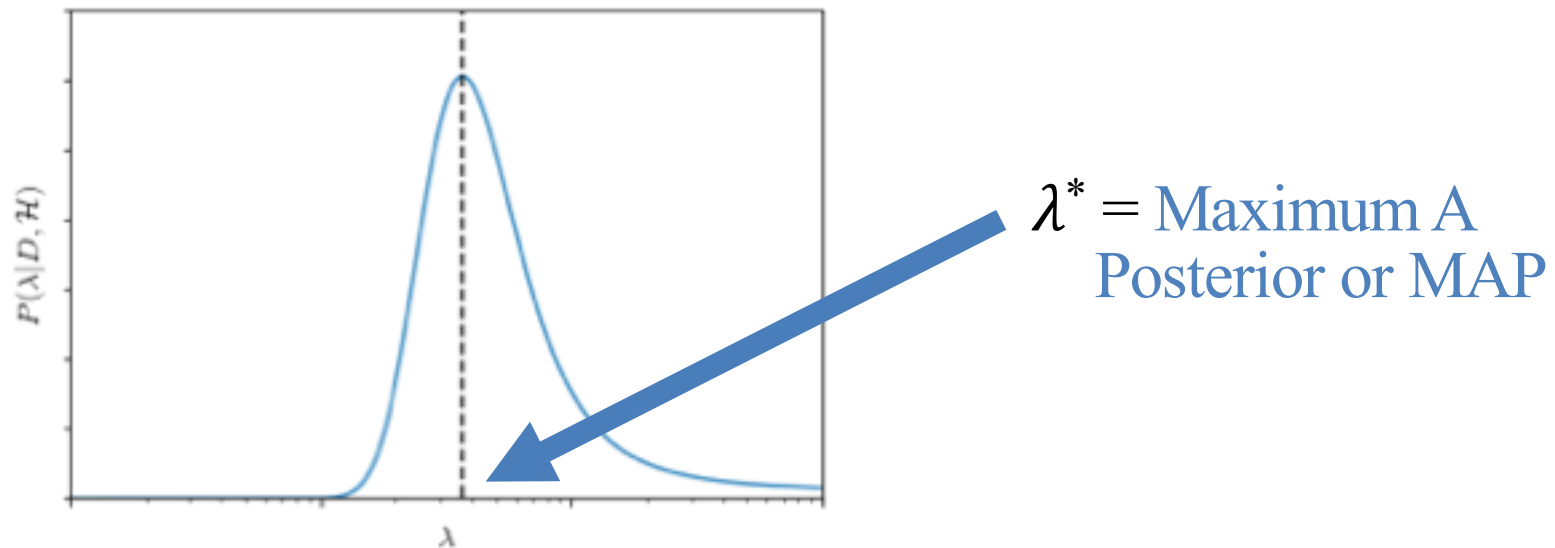


Figure 2.2 *Hypothetical density for which the 95% central interval and 95% highest posterior density region dramatically differ: (a) central posterior interval, (b) highest posterior density region.*

Posterior  $p(\lambda|D, \mathcal{H})$  contains all information on  $\lambda$



If  $p(\lambda) \propto \text{constant}$  (uniform prior)  $\rightarrow \lambda^* = \text{maximum likelihood estimator (MLE)}$

$$\left. \frac{d}{d\lambda} p(\lambda|D, \mathcal{H}) \right|_{\lambda=\lambda^*} = 0 \quad \text{Approximate } p(\lambda|D) \text{ as a Gaussian around } \lambda^*$$

- Error estimate:  $-\left. \frac{d^2}{d\lambda^2} \ln p(\lambda|D, \mathcal{H}) \right|_{\lambda=\lambda^*} = \frac{1}{\sigma_\lambda^2}$
- Laplace approximation:  $\lambda = \lambda^* \pm \sigma_\lambda$

# Posterior predictive distribution

- Predicting future observation conditional on current data  $y$  and model posterior: we marginalize over all models at fixed current data  $y$
- Main idea: future data are drawn from a model but with some scatter

$$\begin{aligned} p(\tilde{y}|y) &= \int p(\tilde{y}|\theta)p(\theta|y)d\theta \\ &\propto \int \exp\left(-\frac{1}{2\sigma^2}(\tilde{y} - \theta)^2\right) \exp\left(-\frac{1}{2\tau_1^2}(\theta - \mu_1)^2\right) d\theta. \end{aligned}$$

$$E(\tilde{y}|y) = E(E(\tilde{y}|\theta, y)|y) = E(\theta|y) = \mu_1,$$

and

$$\begin{aligned} \text{var}(\tilde{y}|y) &= E(\text{var}(\tilde{y}|\theta, y)|y) + \text{var}(E(\tilde{y}|\theta, y)|y) \\ &= E(\sigma^2|y) + \text{var}(\theta|y) \\ &= \sigma^2 + \tau_1^2. \end{aligned}$$

Two sources of uncertainty!

# Modern statistical methods (Bayesian or not)

Gelman et al., *Bayesian Data Analysis*, 3<sup>rd</sup> edition

- a willingness to use many parameters
- hierarchical structuring of models, which is the essential tool for achieving partial pooling of estimates and compromising in a scientific way between alternative sources of information
- model checking—not only by examining the internal goodness of fit of models to observed and possible future data, but also by comparing inferences about estimands and predictions of interest to substantive knowledge
- an emphasis on inference in the form of distributions or at least interval estimates rather than simple point estimates
- the use of simulation as the primary method of computation; the modern computational counterpart to a ‘joint probability distribution’ is a set of randomly drawn values, and a key tool for dealing with missing data is the method of multiple imputation (computation and multiple imputation are discussed in more detail in later chapters)
- the use of probability models as tools for understanding and possibly improving data-analytic techniques that may not explicitly invoke a Bayesian model
- the importance of including in the analysis as much background information as possible, so as to approximate the goal that data can be viewed as a random sample, conditional on all the variables in the model
- the importance of designing studies to have the property that inferences for estimands of interest will be robust to model assumptions.

# INTRODUCTION TO MODELING OF DATA

- We are given  $N$  number of data measurements  $(x_i, y_i)$
- Each measurement comes with an error estimate  $\sigma_i$
- We have a parametrized model for the data  $y = y(x_i)$
- We think the error probability is Gaussian and the measurements are uncorrelated:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma_i^2}} e^{-\frac{(y(x_i) - y_i)^2}{2\sigma_i^2}}$$

$$p(\vec{y}) = \prod_i p(y_i)$$



We can parametrize the model in terms of  $M$  free parameters

$$y(x_i|a_1, a_2, a_3, \dots, a_M)$$

Bayesian formalism gives us the full posterior information on the parameters of the model

$$p(\vec{y}|\vec{a}) = \prod_i p(y_i|\vec{a}) = \mathcal{L}(\vec{a})$$

$$p(a_1, \dots, a_M|\vec{y}) = \frac{\prod_i p(y_i|\vec{a})p(\vec{a})}{p(y_i)}$$

We can assume a flat prior  $p(a_1, a_2, a_3, \dots, a_M) = \text{constant}$

In this case posterior proportional to likelihood

Normalization (evidence, marginal)  $p(y_i)$  not needed if we just need relative posterior density

# Maximum likelihood estimator (MLE)

- Instead of the full posterior we can ask what is the best fit value of parameters  $a_1, a_2, a_3, \dots, a_M$
- We can define this in different ways: **mean, median, mode**
- Choosing the mode (peak posterior or peak likelihood) means we want to **maximize the likelihood: maximum likelihood estimator** (or MAP for non-uniform prior)

$$\text{MLE : } \frac{\partial \mathcal{L}}{\partial \vec{a}} = 0 \quad \text{or} \quad \frac{\partial \ln \mathcal{L}}{\partial \vec{a}} = 0$$



## Maximum likelihood estimator

$$-2\ln\mathcal{L} = \underbrace{\sum_i \left\{ \frac{(y_i - y(x_i|a_1, \dots, a_M))^2}{\sigma_i^2} + \ln\sigma_i \right\}}_{\chi^2}$$

Since  $\sigma$  does not depend on  $a_i$ , MLE means minimizing  $\chi^2$

$$\frac{\partial\chi^2}{\partial a_k} = 0 \quad \rightarrow \quad \sum_i \frac{y_i - y(x_i)}{\sigma_i^2} \frac{\partial y(x_i)}{\partial a_k} = 0$$

This is a system of  $M$  nonlinear equations for  $M$  unknowns

# Fitting data to a straight line

Linear Regression  $y(x) = y(x; a, b) = a + bx$

$$\chi^2(a, b) = \sum_{i=1}^N \left( \frac{y_i - a - bx_i}{\sigma_i} \right)^2$$

Measures how well the model agrees with the data

Minimize  $\chi^2$ :


$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^N \frac{y_i - a - bx_i}{\sigma_i^2}$$
$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^N \frac{x_i (y_i - a - bx_i)}{\sigma_i^2}$$

Define:

$$S \equiv \sum_{i=1}^N \frac{1}{\sigma_i^2} \quad S_x \equiv \sum_{i=1}^N \frac{x_i}{\sigma_i^2} \quad S_y \equiv \sum_{i=1}^N \frac{y_i}{\sigma_i^2}$$

$$S_{xx} \equiv \sum_{i=1}^N \frac{x_i^2}{\sigma_i^2} \quad S_{xy} \equiv \sum_{i=1}^N \frac{x_i y_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial a} = -2 \sum_{i=1}^N \frac{y_i - a - bx_i}{\sigma_i^2}$$

$$0 = \frac{\partial \chi^2}{\partial b} = -2 \sum_{i=1}^N \frac{x_i (y_i - a - bx_i)}{\sigma_i^2}$$


$$aS + bS_x = S_y$$

$$aS_x + bS_{xx} = S_{xy}$$

Matrix Form:

$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

Solve this with linear algebra

$$\begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix} \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} S_y \\ S_{xy} \end{pmatrix}$$

**Solution:** Define  $\Delta \equiv SS_{xx} - (S_x)^2$

$$\hat{a} = \frac{S_{xx}S_y - S_xS_{xy}}{\Delta}$$

$$\hat{b} = \frac{SS_{xy} - S_xS_y}{\Delta}$$

This gives best fit  $\hat{a}$  &  $\hat{b}$

## What about the errors?

- We approximate the log posterior around its peak with a quadratic function
- The posterior is thus approximated as a Gaussian
- This goes under name Laplace approximation
- Note that the errors need to be described as a matrix

$$-2 \cdot \ln p(a, b | y_i) = -2 \cdot \ln \mathcal{L}(a, b)$$

Taylor expansion around the peak ( $\hat{a}$  &  $\hat{b}$ )

Let  $a = x_1, b = x_2$

$$-2 \cdot \ln \mathcal{L}(x_1, x_2) = -2 \cdot \ln \mathcal{L}(\hat{x}_1, \hat{x}_2) - 2 \cdot \frac{1}{2} \sum_{i,j=1,2} \left. \frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j} \right|_{x_i=\hat{x}_i} \Delta x_i \Delta x_j$$

where  $\Delta x_i = x_i - \hat{x}_i$

$$-\frac{1}{2} \sum_{ij} \Delta x_i C_{ij}^{-1} \Delta x_j$$

Note:  $\langle \Delta x_i \Delta x_j \rangle = C_{ij}$

Gaussian posterior approximation

$$-\frac{\partial^2 \ln \mathcal{L}}{\partial x_i \partial x_j} \equiv C_{ij}^{-1} \quad (C^{-1} = \alpha \text{ is called precision matrix.})$$

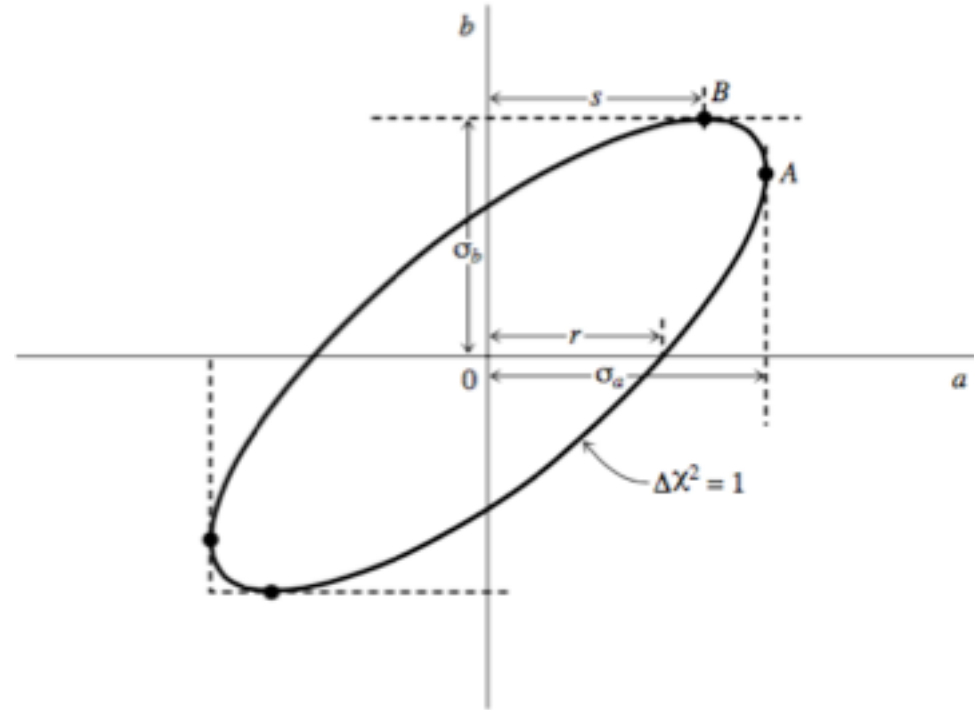
$$\mathcal{L} \propto e^{-\frac{1}{2} \sum_{ij} \Delta x_i C_{ij}^{-1} \Delta x_j}$$

$$-2 \cdot \ln \mathcal{L} = \chi^2$$

$$\frac{\partial^2 \chi^2}{\partial a^2} = 2 \sum_i \frac{1}{\sigma_i^2} = 2S$$

$$\frac{\partial^2 \chi^2}{\partial b^2} = 2 \sum_i \frac{x_i^2}{\sigma_i^2} = 2S_{xx}$$

$$\frac{\partial^2 \chi^2}{\partial a \partial b} = 2 \sum_i \frac{x_i}{\sigma_i^2} = 2S_x$$



$$C^{-1} = \begin{pmatrix} S & S_x \\ S_x & S_{xx} \end{pmatrix}$$

$$C = \frac{1}{\Delta} \begin{pmatrix} S_{xx} & -S_x \\ -S_x & S \end{pmatrix}$$

Define  $\Delta \equiv SS_{xx} - (S_x)^2$

$$\sigma_a^2 = S_{xx}/\Delta \quad \sigma_b^2 = S/\Delta$$



## Bayesian View

- The posterior distribution  $p(a,b|y_i)$  is described as a 2-d  $C^{-1}$  ellipse in  $(a,b)$  plane
- At any fixed value of  $a$  (or  $b$ ) the posterior of  $b$  (or  $a$ ) is a gaussian with variance  $[C^{-1}_{bb(aa)}]^{-1}$
- If we want to know the error on  $b$  (or  $a$ ) independent of  $a$  (or  $b$ ) we need to marginalize over  $a$  (or  $b$ )
- This marginalization can be done analytically, and leads to  $C_{bb(aa)}$  as the variance of  $b$  (or  $a$ )
- This will increase the error:  $C_{bb(aa)} > [C^{-1}_{bb(aa)}]^{-1}$

# Asymptotics theorems

(Le Cam 1953, adopted to Bayesian posteriors)

- Posteriors approach a multi-variate Gaussian in the large  $N$  limit ( $N$ : number of data points):

this is because the 2<sup>nd</sup> order Taylor expansion of  $\ln L$  is more and more accurate in this limit, i.e. we can drop 3<sup>rd</sup> order terms

- The marginalized means approach the true value and the variance approaches the Fisher matrix, defined as ensemble average of precision matrix  $\langle C^{-1} \rangle$
- The likelihood dominates over the prior in large  $N$  limit

# Asymptotics theorems

(Le Cam 1953, adopted to Bayesian posteriors)

- There are caveats when this does not apply, e.g. when data are not informative about a parameter or some linear combination of them, when number of parameters  $M$  is comparable to  $N$ , when posteriors are improper or likelihoods are unbounded... Always exercise care!
- In practice the asymptotic limit is often not achieved for nonlinear models, i.e. we cannot linearize the model across the region of non-zero posterior: this is why we often use sampling to evaluate the posteriors instead of Gaussian
- It is useful to know the existence of this limit, but since we cannot know ahead of time whether we are in this limit or not in practice we cannot assume it: we will be doing full Bayesian posteriors in this course, but we will also compare to the gaussian limit

# Multivariate linear least squares

- We can generalize the model to a generic functional form

$$y_i = a_0 X_0(x_i) + a_1 X_1(x_i) + \dots + a_{M-1} X_{M-1}(x_i)$$

- The problem is linear in  $a_j$  and can be nonlinear in  $x_i$ ,

e.g.  $X_j(x_i) = x_i^j$

$$\chi^2 = \sum_{i=0}^{N-1} \left[ \frac{y_i - \sum_{k=0}^{M-1} a_k X_k(x_i)}{\sigma_i} \right]^2$$

- We can define design matrix  $A_{ij} = X_j(x_i)/\sigma_i$  and

- $b_i = y_i / \sigma_i$

$$\chi^2 = |\mathbf{A} \cdot \mathbf{a} - \mathbf{b}|^2$$

# Design matrix

$$\begin{array}{c}
 \longleftarrow \text{basis functions} \longrightarrow \\
 X_0(\quad) \quad X_1(\quad) \quad \dots \quad X_{M-1}(\quad) \\
 \\
 \begin{array}{c}
 \uparrow \\
 x_0 \\
 \\
 x_1 \\
 \\
 \vdots \\
 \\
 x_{N-1} \\
 \downarrow
 \end{array}
 \begin{pmatrix}
 \frac{X_0(x_0)}{\sigma_0} & \frac{X_1(x_0)}{\sigma_0} & \dots & \frac{X_{M-1}(x_0)}{\sigma_0} \\
 \frac{X_0(x_1)}{\sigma_1} & \frac{X_1(x_1)}{\sigma_1} & \dots & \frac{X_{M-1}(x_1)}{\sigma_1} \\
 \vdots & \vdots & & \vdots \\
 \vdots & \vdots & & \vdots \\
 \frac{X_0(x_{N-1})}{\sigma_{N-1}} & \frac{X_1(x_{N-1})}{\sigma_{N-1}} & \dots & \frac{X_{M-1}(x_{N-1})}{\sigma_{N-1}}
 \end{pmatrix}
 \end{array}$$

*Credit: NR, Press et al.* **20**

# Solution by normal equations

$$0 = \sum_{i=0}^{N-1} \frac{1}{\sigma_i^2} \left[ y_i - \sum_{j=0}^{M-1} a_j X_j(x_i) \right] X_k(x_i) \quad k = 0, \dots, M-1 \quad (15.4.6)$$

Interchanging the order of summations, we can write (15.4.6) as the matrix equation

$$\sum_{j=0}^{M-1} \alpha_{kj} a_j = \beta_k \quad (15.4.7)$$

where

$$\alpha_{kj} = \sum_{i=0}^{N-1} \frac{X_j(x_i) X_k(x_i)}{\sigma_i^2} \quad \text{or, equivalently,} \quad \boldsymbol{\alpha} = \mathbf{A}^T \cdot \mathbf{A} \quad (15.4.8)$$

an  $M \times M$  matrix, and

$$\beta_k = \sum_{i=0}^{N-1} \frac{y_i X_k(x_i)}{\sigma_i^2} \quad \text{or, equivalently,} \quad \boldsymbol{\beta} = \mathbf{A}^T \cdot \mathbf{b} \quad (15.4.9)$$

$$\boldsymbol{\alpha} \cdot \mathbf{a} = \boldsymbol{\beta} \quad \text{or as} \quad (\mathbf{A}^T \cdot \mathbf{A}) \cdot \mathbf{a} = \mathbf{A}^T \cdot \mathbf{b} \quad (15.4.10)$$

To solve the normal equations to obtain best fit values and the precision matrix we need to learn **linear algebra** numerical methods: topic of next lecture

# Gaussian posterior

$$P(\delta \mathbf{a}) da_0 \dots da_{M-1} = \text{const.} \times \exp\left(-\frac{1}{2} \delta \mathbf{a} \cdot \boldsymbol{\alpha} \cdot \delta \mathbf{a}\right) da_0 \dots da_{M-1}$$

## Marginalization over nuisance parameters

- If we want to know the error on  $j$ -th parameter we need to marginalize over all other parameters
- In analogy to 2-d case this leads to  $\sigma_j^2 = C_{jj}$
- So we need to invert the precision matrix  $a = C^{-1}$
- Analytic marginalization is only possible for a multi-variate Gaussian distribution: a great advantage of using a Gaussian
- If the posterior is not Gaussian it may be made more Gaussian by a nonlinear transformation of the variable



Show

$$\int da \cdot e^{-\frac{1}{2} \left[ (a-\hat{a})^2 C_{aa}^{-1} + (a-\hat{a})(b-\hat{b}) C_{ab}^{-1} + (b-\hat{b})^2 C_{bb}^{-1} \right]} \propto e^{-\frac{1}{2} \frac{(b-\hat{b})^2}{C_{bb}}}$$

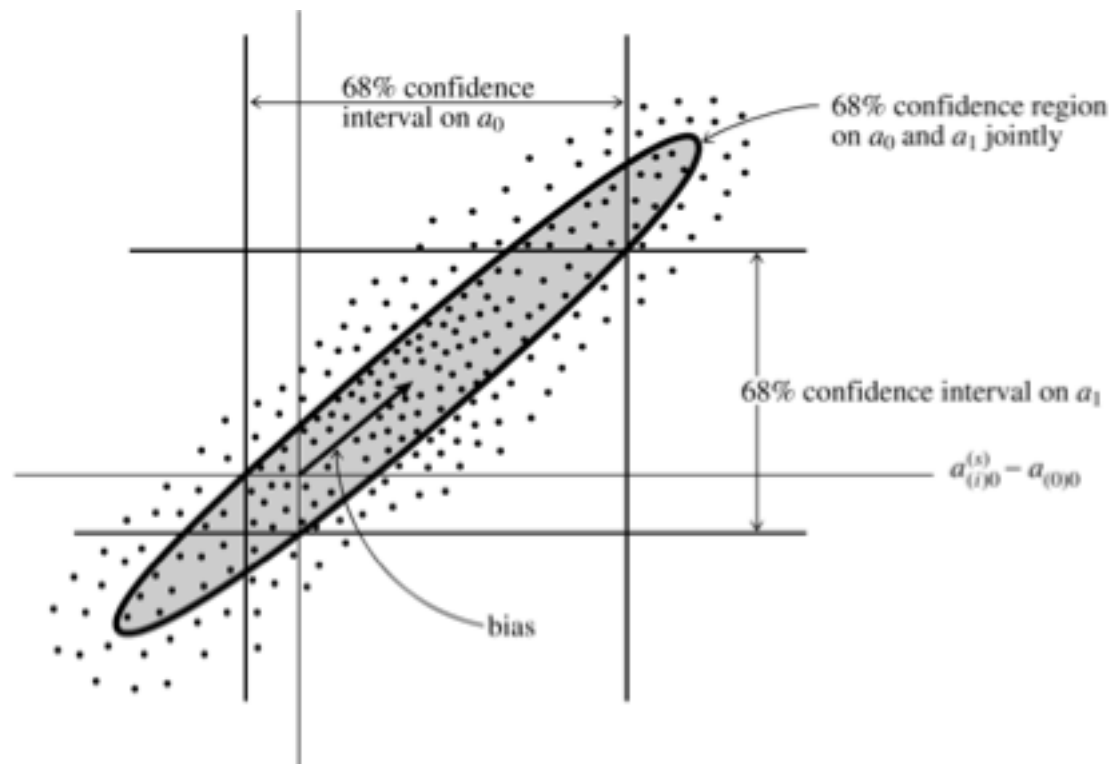
(Complete the square in  $a$ )

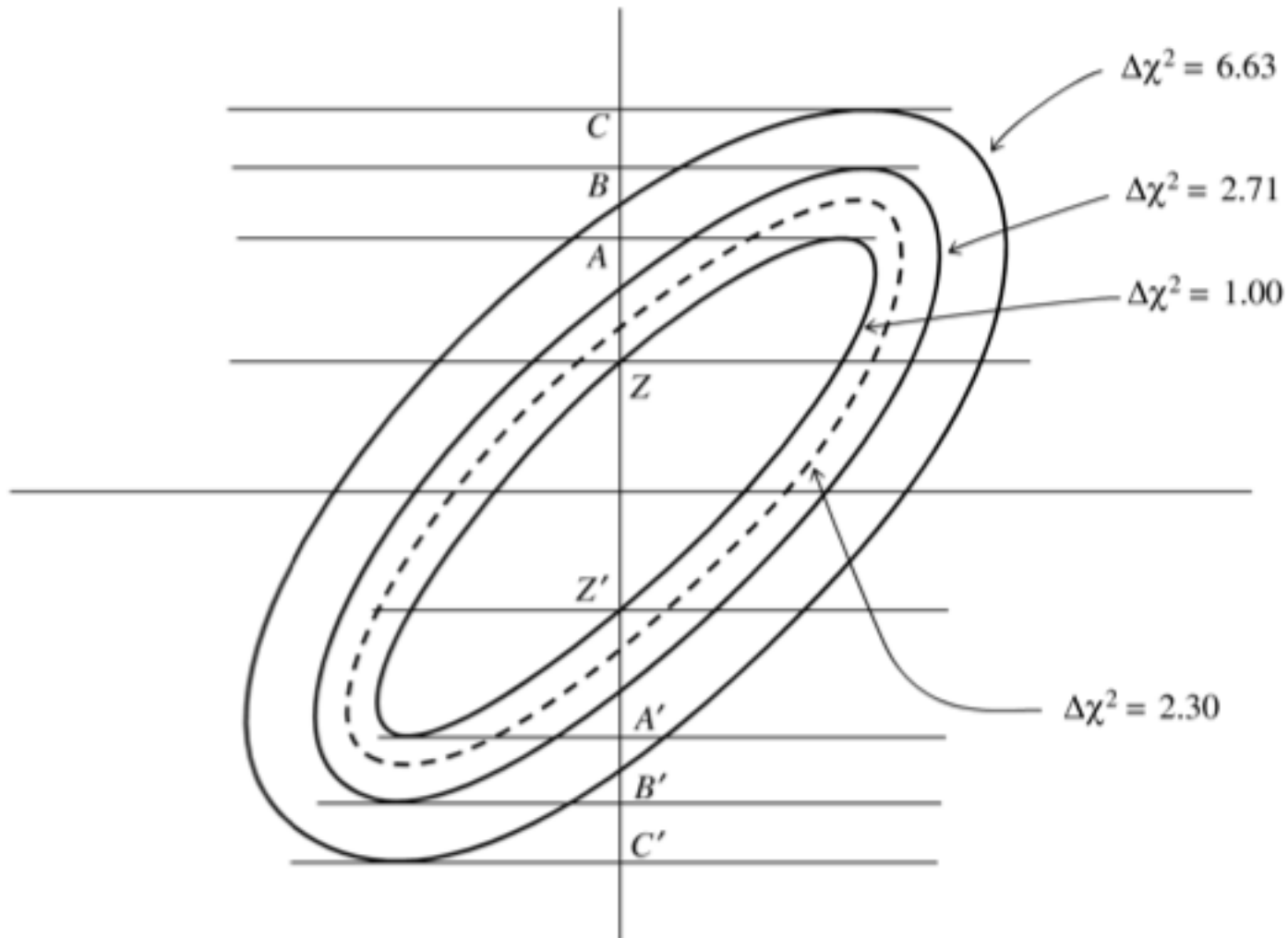
# What about multi-dimensional projections?

- Suppose we are interested in  $\nu$  components of  $a$ , marginalizing over remaining  $M - \nu$  components.
- We take the components of  $C$  corresponding to  $\nu$  parameters to create  $\nu \times \nu$  matrix  $C_{\text{proj}}$
- Invert the matrix to get precision matrix  $C_{\text{proj}}^{-1}$
- Posterior distribution is proportional to  
 $\exp(-\delta a_{\text{proj}}^T C_{\text{proj}}^{-1} \delta a_{\text{proj}}/2),$   
which is distributed as  $\exp(-\Delta\chi^2/2),$   
i.e.  $\chi^2$  with  $\nu$  degrees of freedom

# Credible intervals under Gaussian posterior approx.

- We like to quote posteriors in terms of X% credible intervals
- For Gaussian likelihoods most compact posteriors correspond to a constant change  $\Delta\chi^2$  relative to MAP/MLE
- The intervals depend on the dimension: example for X=68





**Figure 15.6.4.** Confidence region ellipses corresponding to values of chi-square larger than the fitted minimum. The solid curves, with  $\Delta\chi^2 = 1.00, 2.71, 6.63$ , project onto one-dimensional intervals  $AA', BB', CC'$ . These intervals — not the ellipses themselves — contain 68.3%, 90%, and 99% of normally distributed data. The ellipse that contains 68.3% of normally distributed data is shown dashed and has  $\Delta\chi^2 = 2.30$ . For additional numerical values, see the table on p. 815.

$\Delta\chi^2$ as a Function of Confidence Level $p$ and Number of Parameters of Interest $\nu$						
$p$	$\nu$					
	1	2	3	4	5	6
68.27%	1.00	2.30	3.53	4.72	5.89	7.04
90%	2.71	4.61	6.25	7.78	9.24	10.6
95.45%	4.00	6.18	8.02	9.72	11.3	12.8
99%	6.63	9.21	11.3	13.3	15.1	16.8
99.73%	9.00	11.8	14.2	16.3	18.2	20.1
99.99%	15.1	18.4	21.1	23.5	25.7	27.9

We rarely go above  $\nu = 2$  dimensions in projections  
(difficult to visualize)

# Introduction to Machine Learning

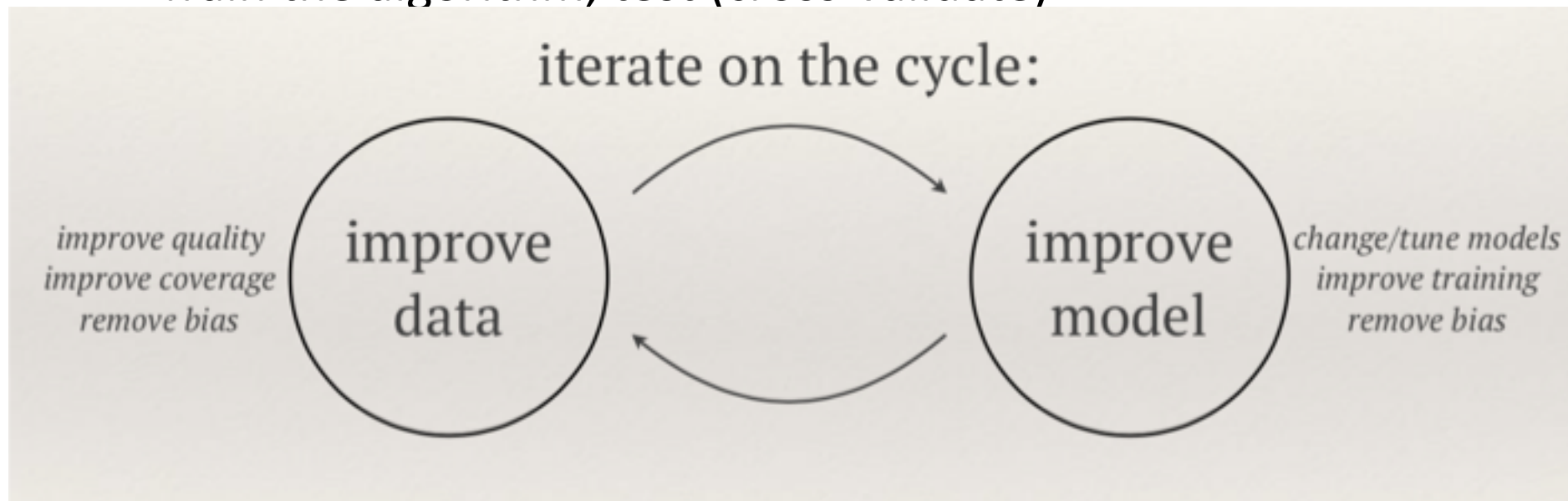
- From some input  $x$ , output can be:
- Summary  $z$ : unsupervised learning (descriptive, hindsight)
- Prediction  $y$ : supervised learning (predictive, insight)
- Action  $a$  to maximize reward  $r$ : reinforcement learning (prescriptive, foresight)
- value/difficulty (subjective and controversial!)
- Supervised learning: classification and regression
- Unsupervised learning: dimensionality reduction



Chris Wiggins taxonomy, Gartner/Recht graph 29

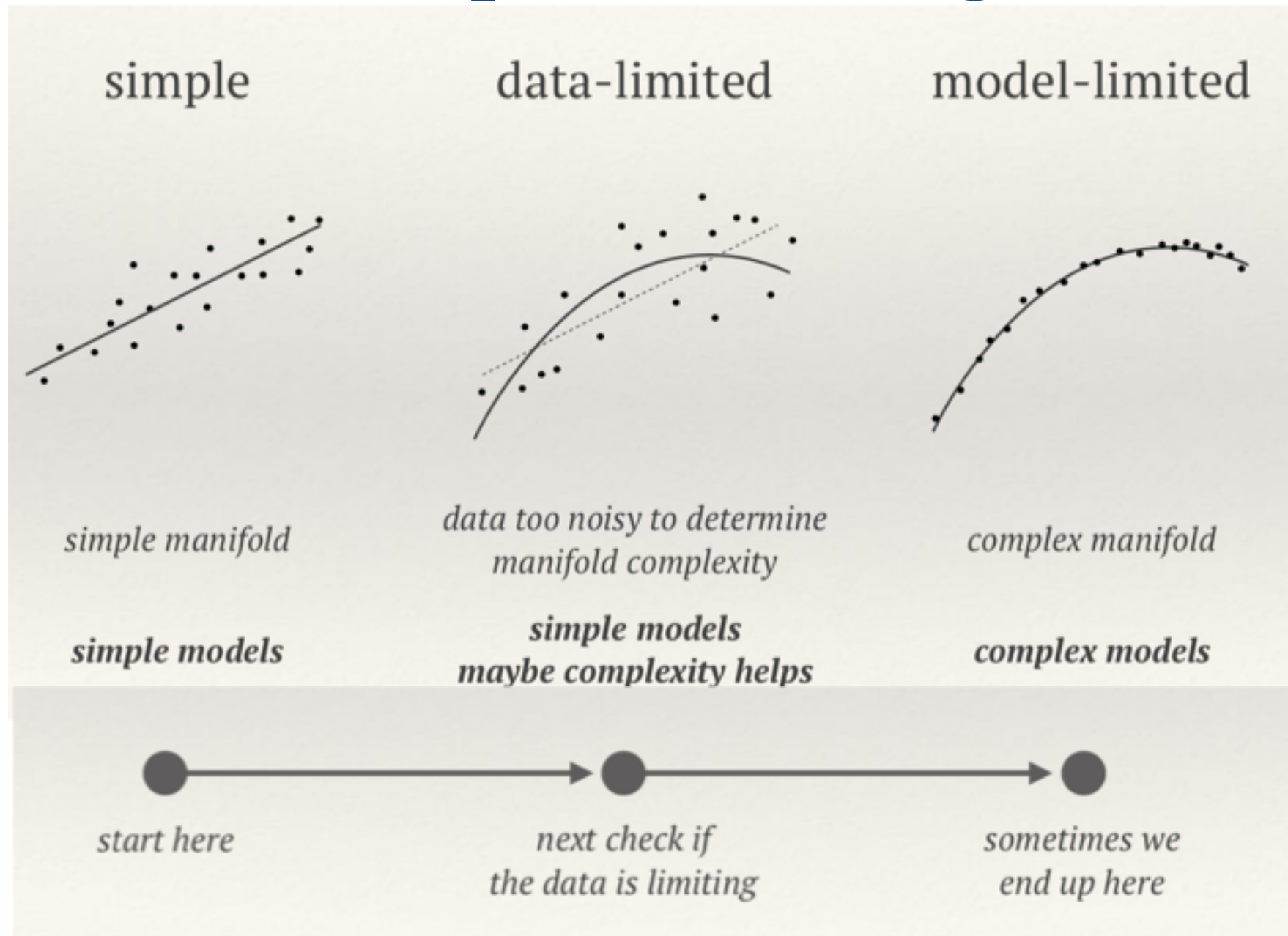
# Supervised Learning (SL)

- Answering a specific question: e.g. regression or classification
- When doing regression this is interpolation/extrapolation
- General approach: frame the problem, collect the data
- Choose the SL algorithm
- Choose objective function (decide what to optimize)
- Train the algorithm, test (cross-validate)





# Classes of problems: regression



# Basic ML procedure

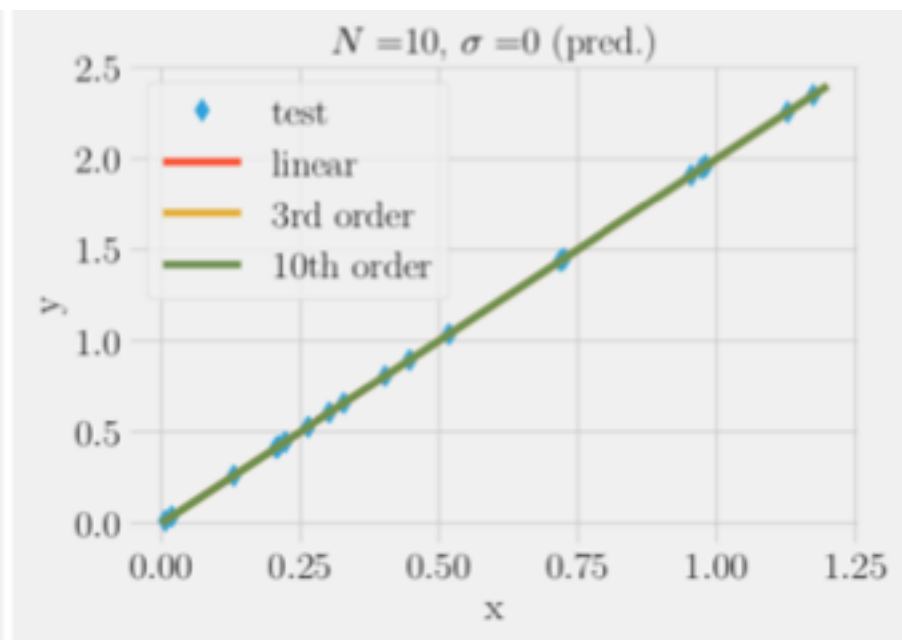
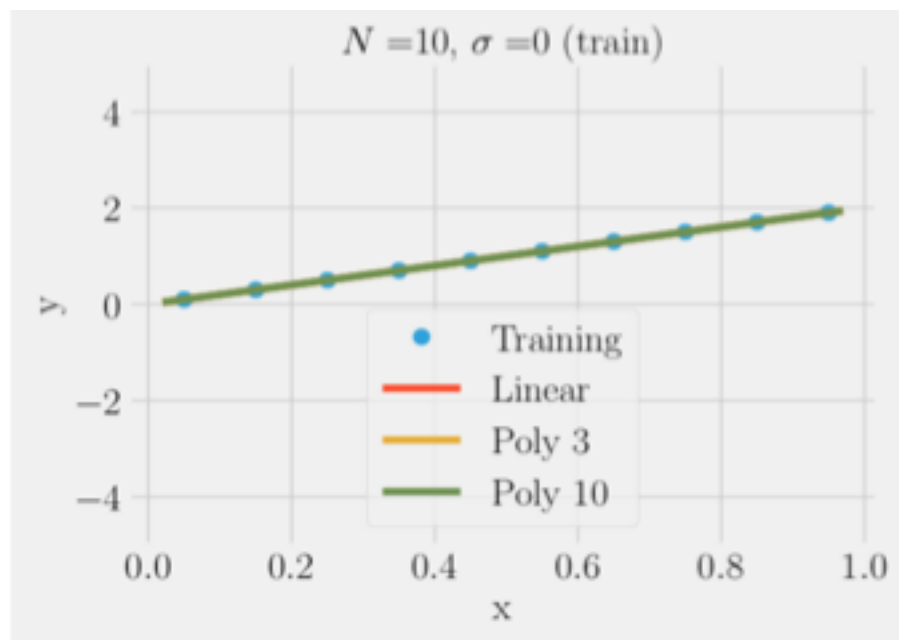
- We have some data  $Y$ , some model  $g(\omega)$  and some cost or loss function  $C(Y, g(\omega))$  that we wish to minimize such that the model  $g(\omega)$  explains the data  $Y$ .

$$\chi^2 = \sum_{i=0}^{N-1} \left[ \frac{y_i - \sum_{k=0}^{M-1} a_k X_k(x_i)}{\sigma_i} \right]^2$$

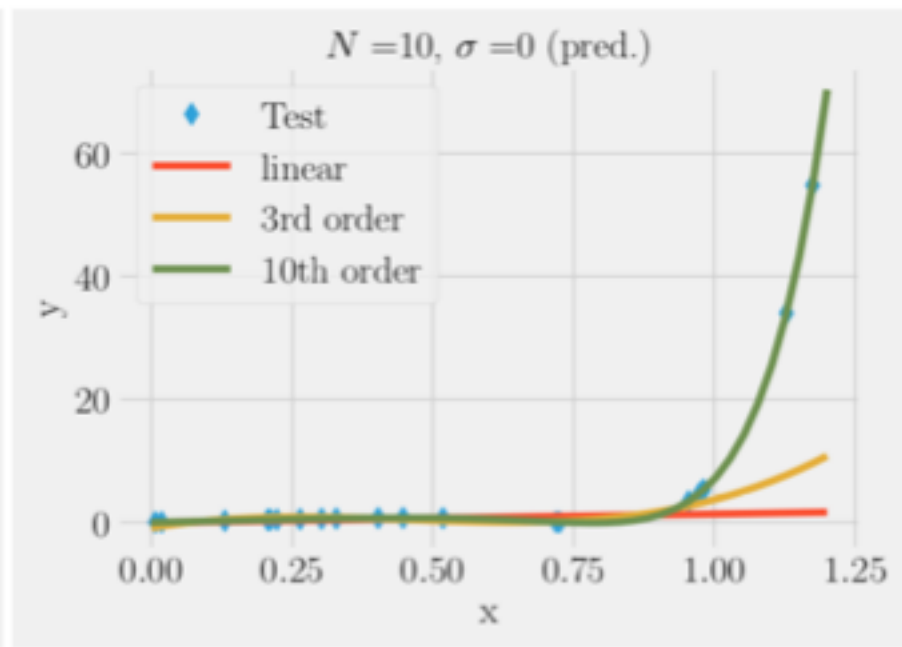
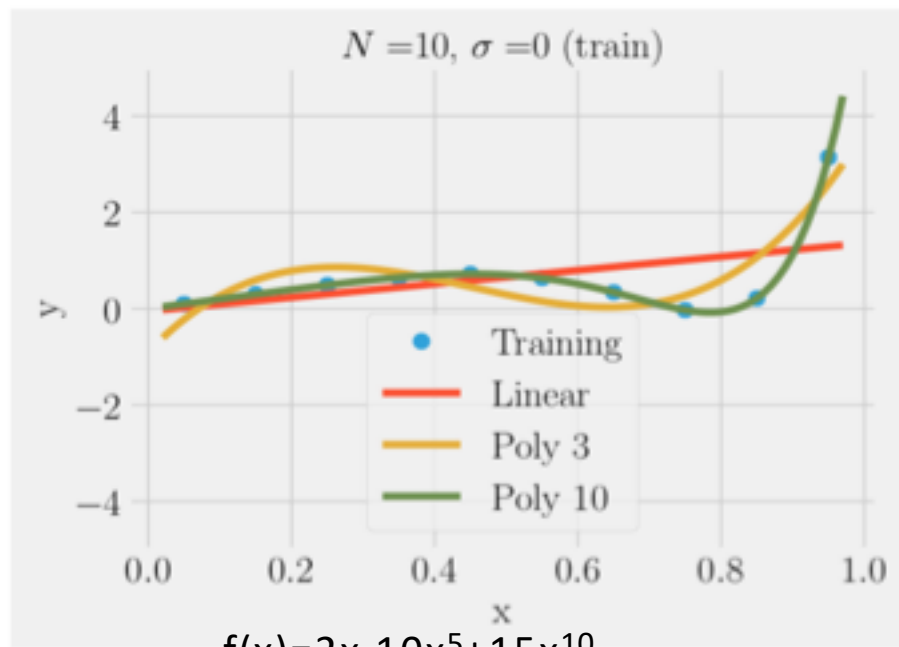
- E.g.  $Y=(x,y)$ ,  $C=\chi^2$ ,  $\omega=a$ ,
- $g=a_0X_0(x_i) + a_1X_1(x_i) + \dots + a_{M-1}X_{M-1}(x_i)$
- In ML we divide data into training data  $Y_{\text{train}}$  (e.g. 90%) and test data  $Y_{\text{test}}$  (e.g. 10%)
- We fit model to the training data: the value of the minimum loss function at  $\omega_{\text{min}}$  is called in-sample error  $E_{\text{in}}=C(Y_{\text{train}},g(\omega_{\text{min}}))$
- we test the results on test data, getting out of sample error  $E_{\text{out}}=C(Y_{\text{test}},g(\omega_{\text{min}}))>E_{\text{in}}$
- This is called cross-validation technique
- If we have different models then test data are called validation data while test data are used to test different models, each trained on training data (3 way split, e.g. 60%, 30%, 10%)

# Data analysis versus machine learning: fitting versus predicting

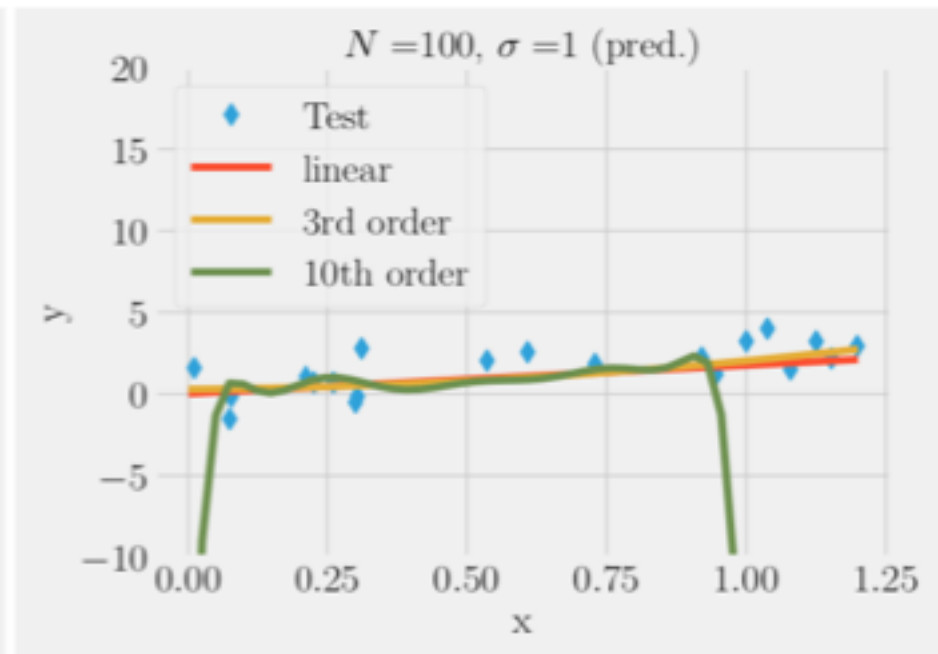
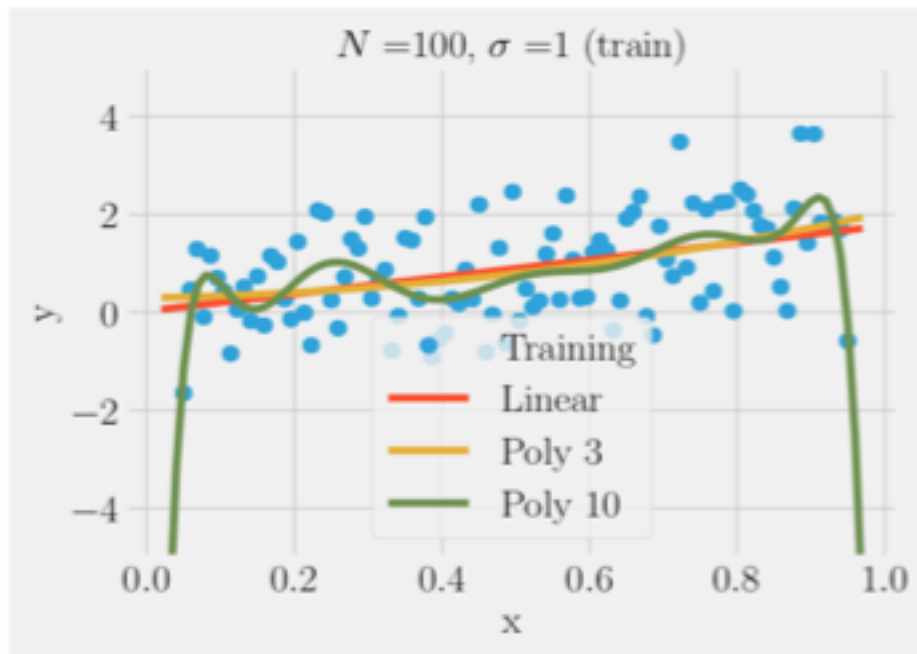
- Data analysis: fitting existing data to obtain model parameters
- ML: use model derived from existing data to predict something (regression, classification) for new data
- Example: polynomial regression. This will be HW 4 problem
- We can fit the training data to a simple model or complex model
- In the absence of noise complex model always better
- In the presence of noise complex model often worse
- A proper Bayesian predicting procedure averages over the uncertainties of model parameters: this is often ignored in ML, or done with bootstrap methods (bagging, to be discussed later)



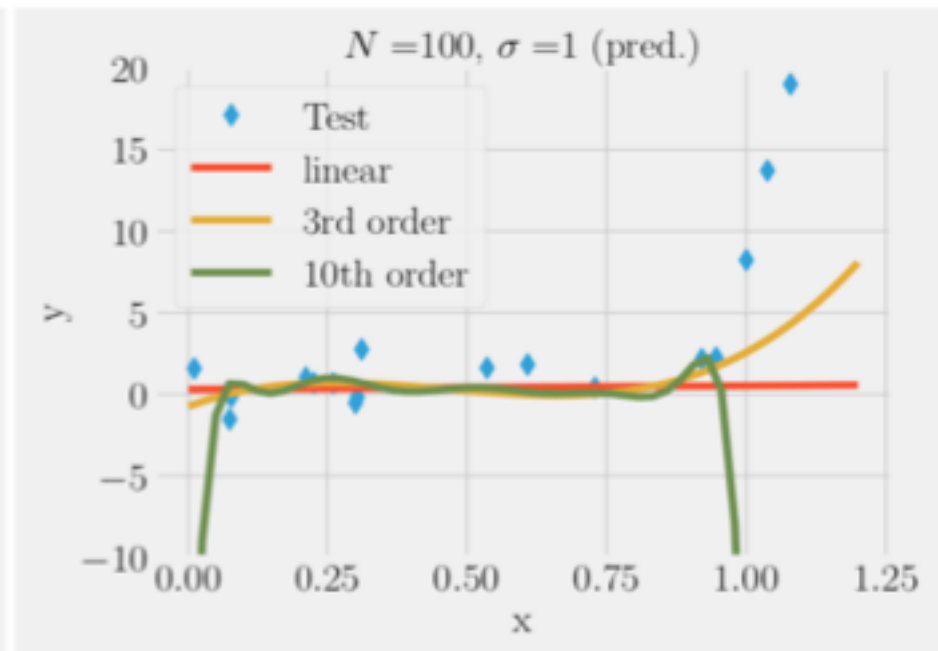
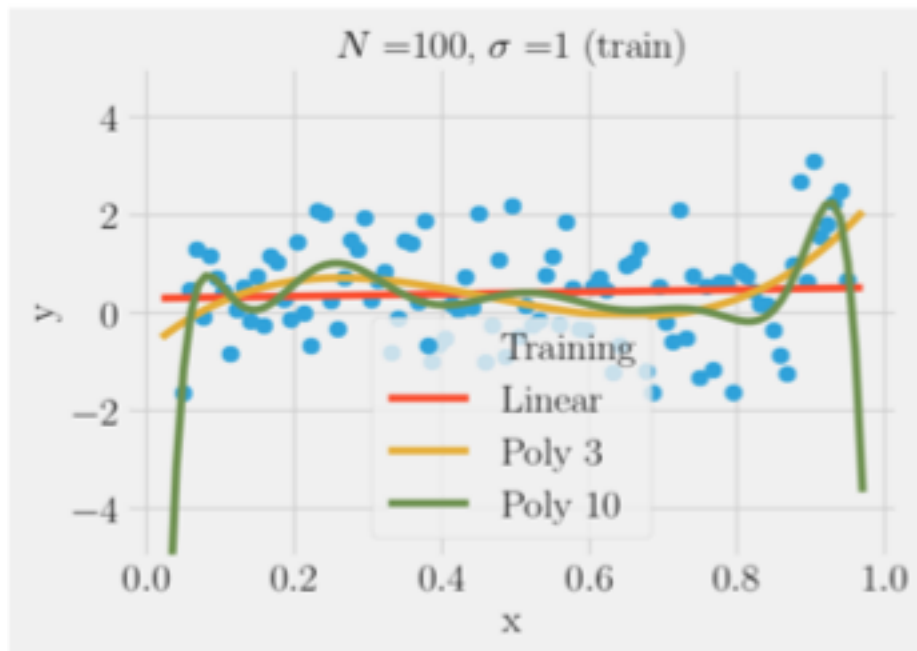
$f(x)=2x$ , no noise



$f(x)=2x-10x^5+15x^{10}$



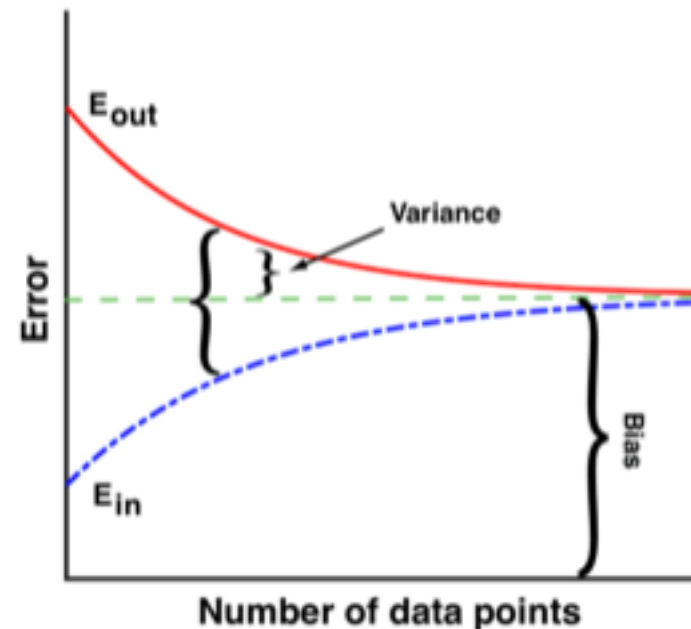
Over-fitting noise with too complex models (bias-variance trade-off)



35

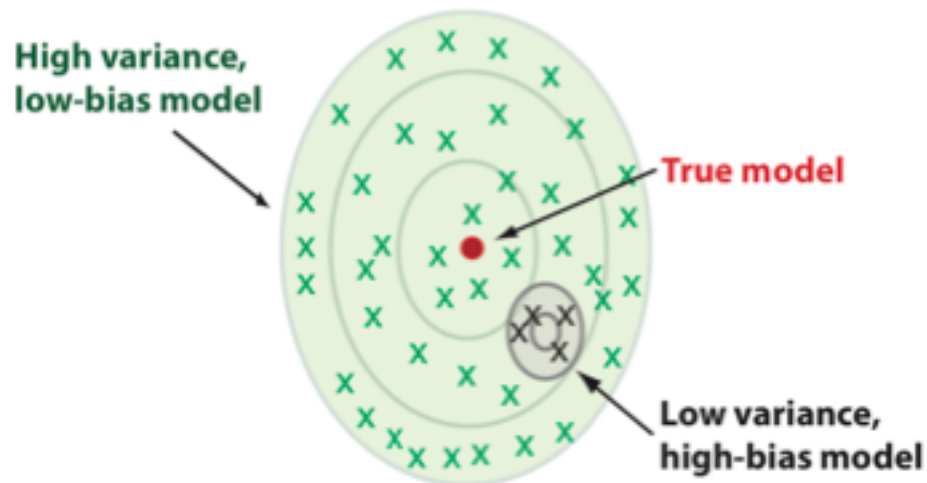
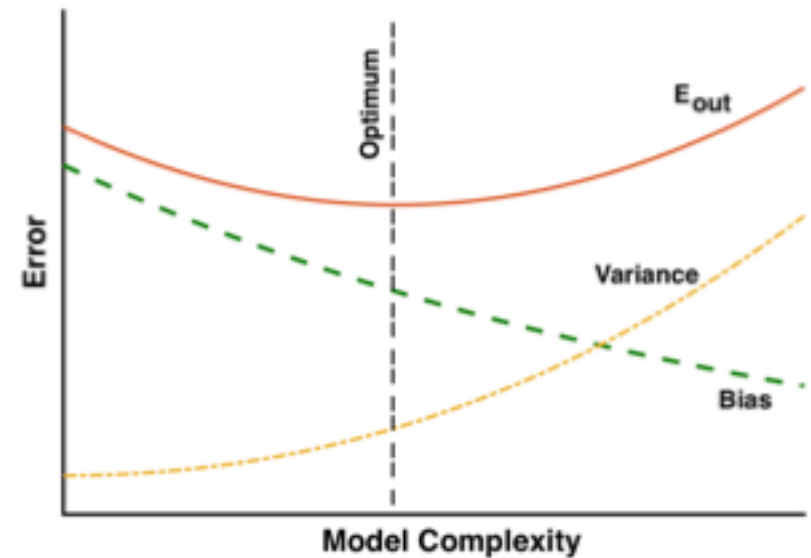
# Statistical learning theory

- We have data and we can change the number of data points
- we have models and we can change complexity (number of model parameters in simple versions)
- Trade-off at fixed model complexity:
- small data size suffers from a large variance
- Large data size suffers from model bias
- Variance quantified by  $E_{in}$  vs  $E_{out}$
- $E_{in}$  and  $E_{out}$  approach bias for large data



# Bias-variance trade-off vs complexity

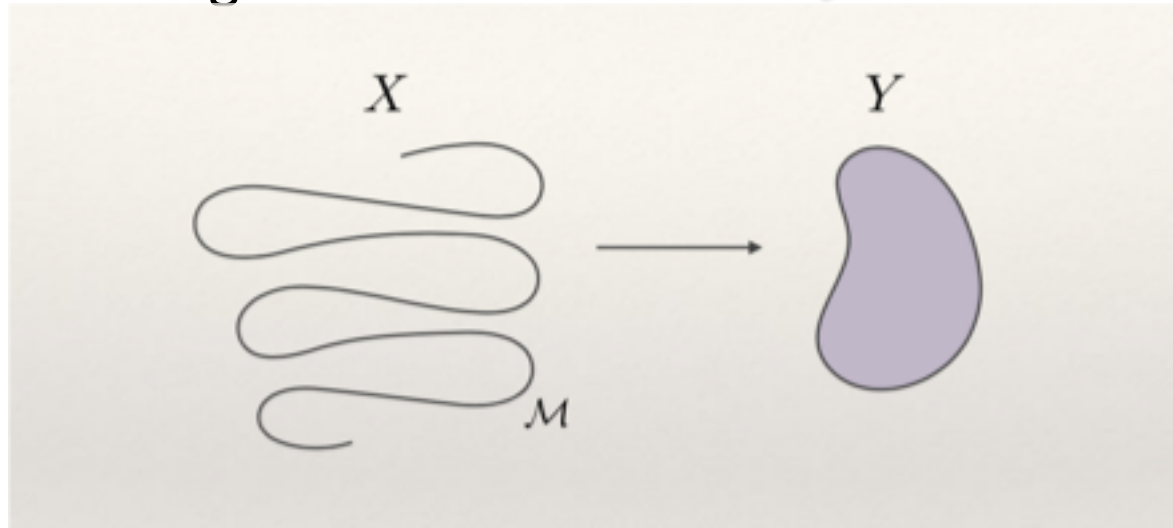
- Low complexity: large bias
- Large complexity: large variance
- Optimum when the two are balanced





# Representational power

- We are learning a manifold  $M$   $f: X \rightarrow Y$



- To learn complex manifolds we need high representational power
- We need a universal approximator with good generalization properties (from in-sample to out of sample, i.e. not over-fitting)
- This is where neural networks excel: they can fit anything (literally, including pure noise), yet can also generalize

38

## Literature

- *Numerical Recipes*, Press et al., Chapter 15  
(<http://apps.nrbook.com/c/index.html>)
- *Bayesian Data Analysis*, Gelman et al. , Chapter 1-4
- *A high bias, low variance introduction to machine learning for physicists*, <https://arxiv.org/pdf/1803.08823.pdf>