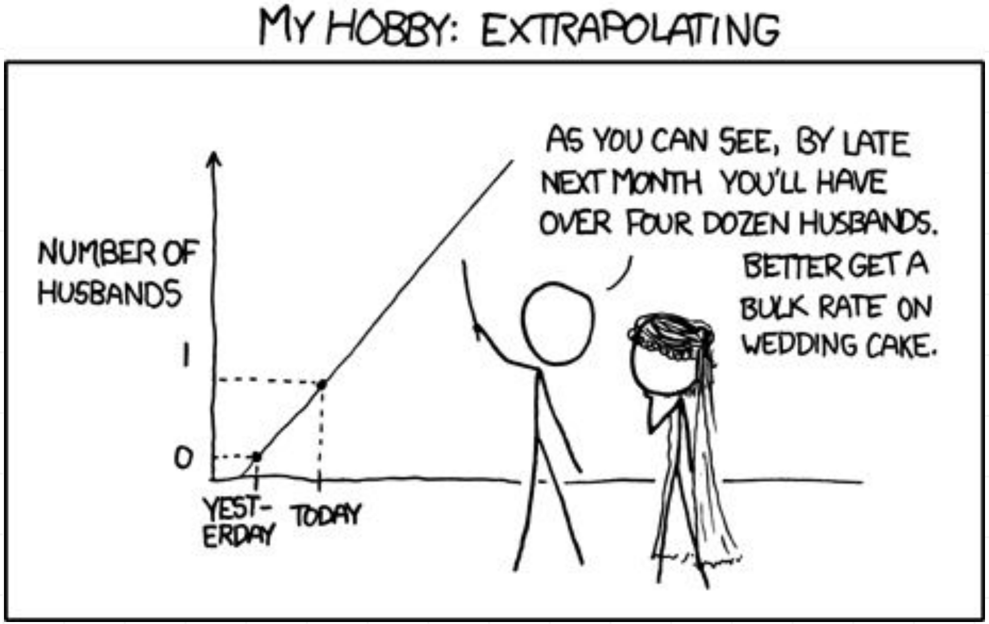
Hold on … SML lecture will be starting soon.

https://xkcd.com/605/



On the topic of extrapolation and train-test mismatch, see https://www.youtube.com/watch?v=es6p6NuxOnY and http://ciml.info/dl/v0_99/ciml-v0_99-ch08.pdf

Plan for Today

ML 101:
Polynomial curve fitting: model, loss/error function, over-fitting, regularisation

Model selection

Probabilities: sum rule, product rule, Bayes theorem
Gaussians - 1D, maximum likelihood estimates (MLE), bias-variance
➔ and how this helps curve-fitting

Gaussians (multidimensional)
various matrix identities, geometric intuitions

Bernoulli, Binomial, Exponential family distributions - will be in assignment 1

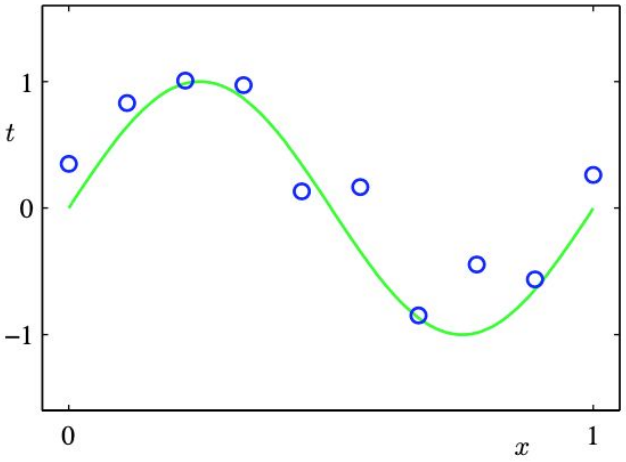
Review: probabilities, derivatives and finding stationary points, eigenvalues and eigenvectors

about the book

Tom Mitchell (1998): a computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E.

some artificial data created from the function

$$\sin(2\pi x) + \text{random noise} \qquad x = 0, \ldots, 1$$



The machine sees:

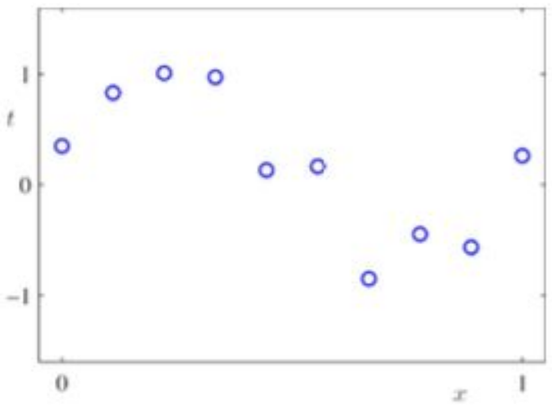
$$N = 10$$

$$\mathbf{x} \equiv (x_1, \ldots, x_N)^T$$

$$\mathbf{t} \equiv (t_1, \ldots, t_N)^T$$

$$x_i \in \mathbb{R} \quad i = 1, \ldots, N$$

$$t_i \in \mathbb{R} \quad i = 1, \ldots, N$$



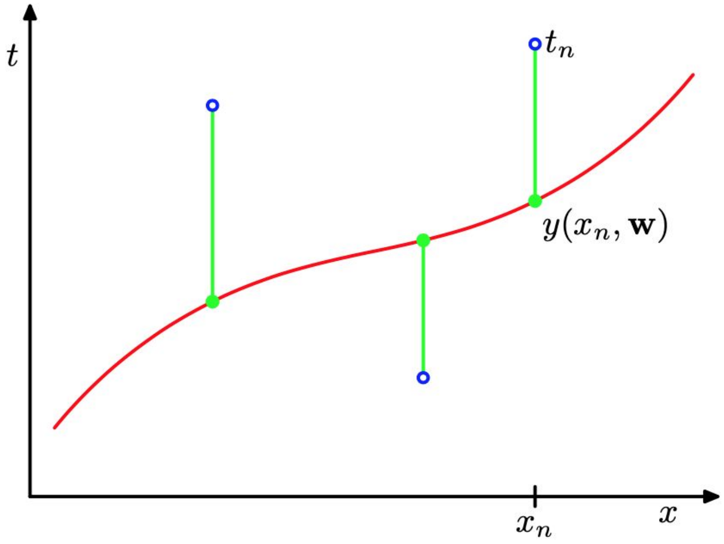
Make a guess, M-th order polynomials

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \qquad (1.1)$$

What is a good “fit”

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \qquad (1.2)$$

Figure 1.3 The error function (1.2) corresponds to (one half of) the sum of the squares of the displacements (shown by the vertical green bars) of each data point from the function $y(x, \mathbf{w})$.



$$y(x, \mathbf{w}) = w_0$$

$$w_0 + w_1 x$$

$$w_0 + w_1 x + w_2 x^2 + w_3 x^3$$

$$w_0 + w_1 x + \cdots + w_8 x^8 + w_9 x^9$$

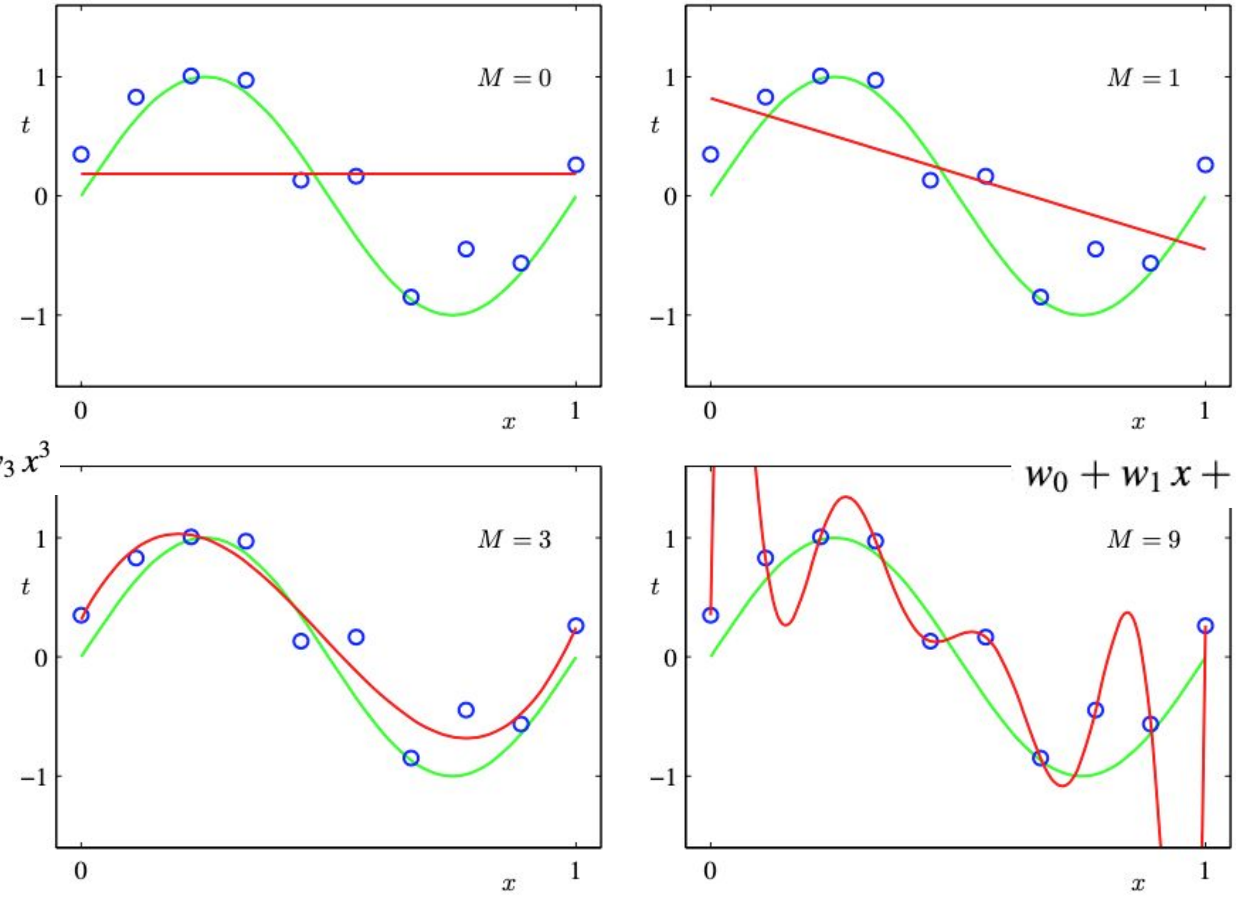


Figure 1.4 Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

Test error and learning curves

Training set: 10 points

Separate test set of 100 points

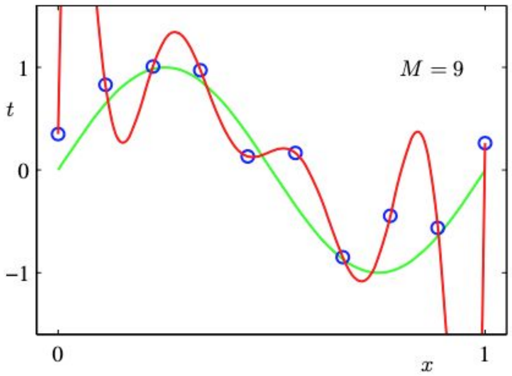
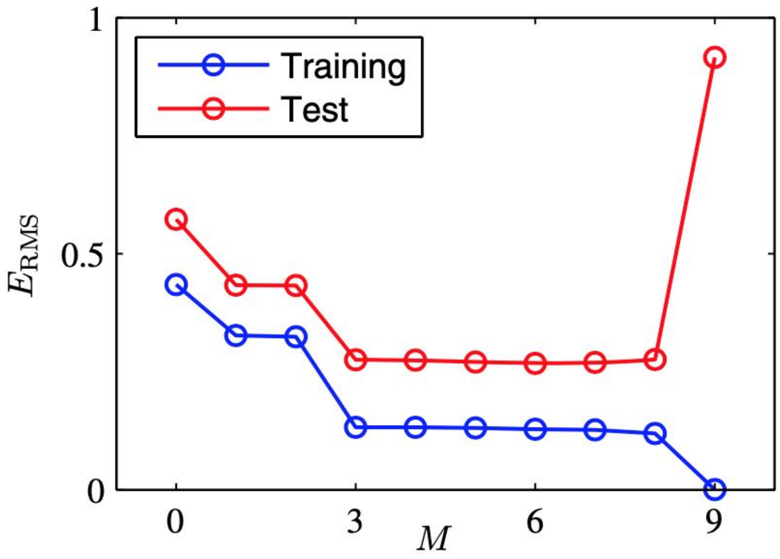


Figure 1.5 Graphs of the root-mean-square error, defined by (1.3), evaluated on the training set and on an independent test set for various values of M.



$$E_{\mathrm{RMS}} = \sqrt{2E(\mathbf{w}^{\star})/N}$$

why?

Expansion of sin(x) contains terms of all orders

Table 1.1 Table of the coefficients $\mathbf{w}^\star$ for polynomials of various order. Observe how the typical magnitude of the coefficients increases dramatically as the order of the polynomial increases.

	$M = 0$	$M = 1$	$M = 6$	$M = 9$
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

Cure 1: More data :)

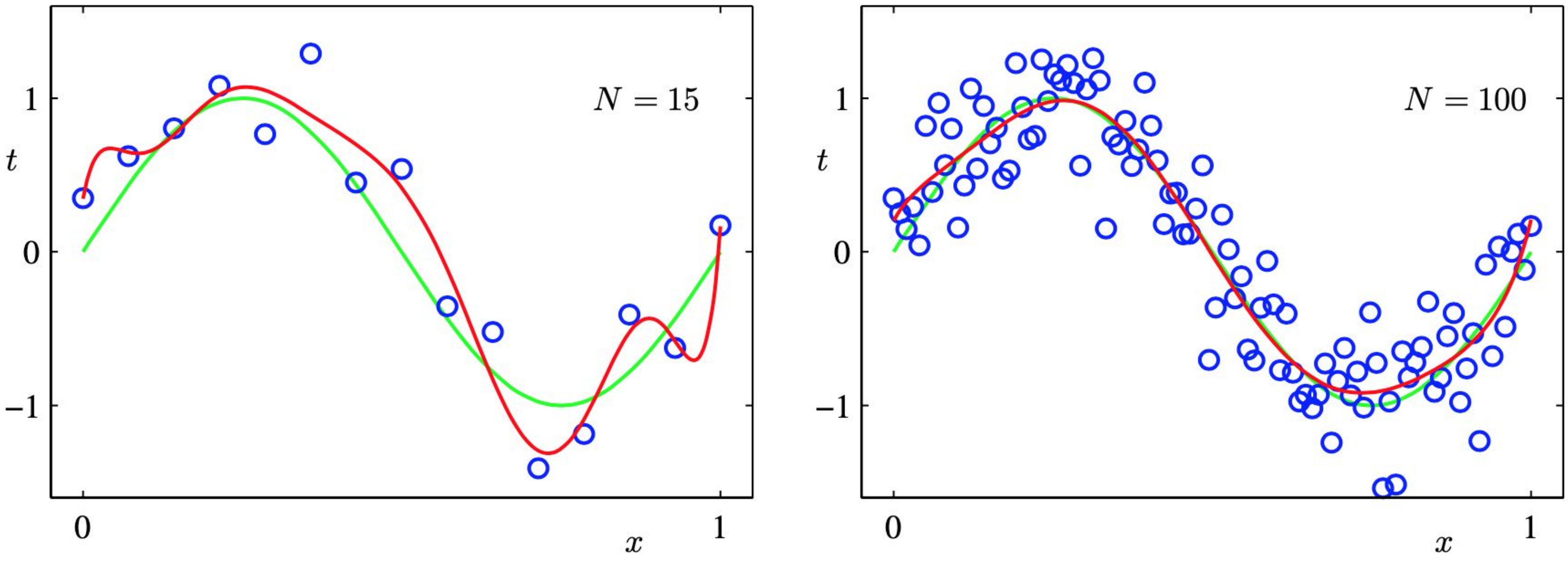


Figure 1.6 Plots of the solutions obtained by minimizing the sum-of-squares error function using the $M = 9$ polynomial for $N = 15$ data points (left plot) and $N = 100$ data points (right plot). We see that increasing the size of the data set reduces the over-fitting problem.

Cure 2: regularisation

Minimize *regularised* error function

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \qquad (1.4)$$

(more in Bayesian regression next week)

$$\widetilde{E}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{\lambda}{2}\|\mathbf{w}\|^2 \tag{1.4}$$

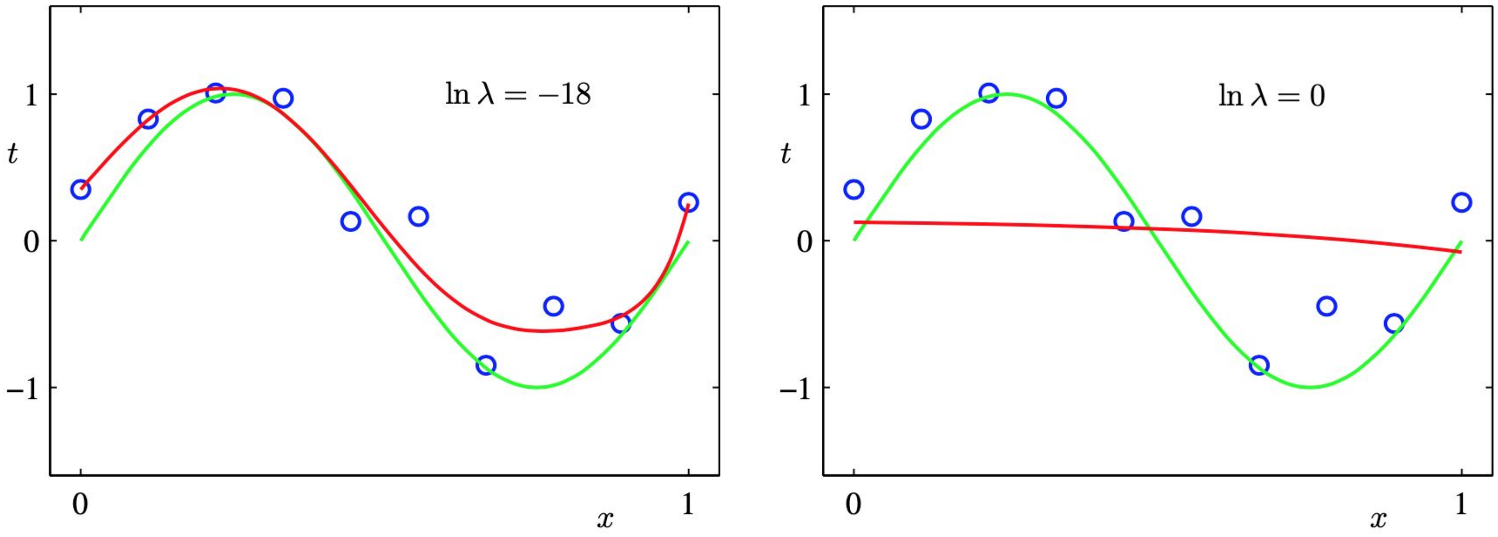
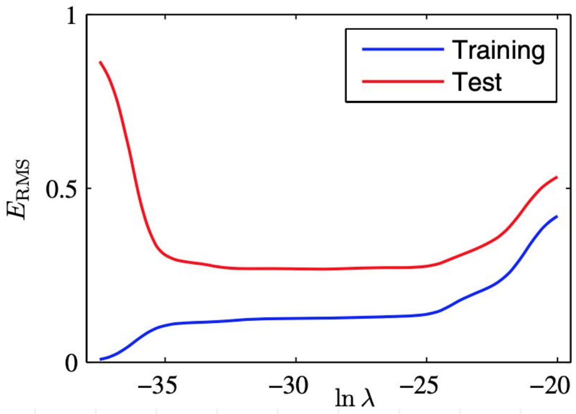


Figure 1.7 Plots of $M = 9$ polynomials fitted to the data set shown in Figure 1.2 using the regularized error function (1.4) for two values of the regularization parameter λ corresponding to $\ln\lambda = -18$ and $\ln\lambda = 0$. The case of no regularizer, i.e., $\lambda = 0$, corresponding to $\ln\lambda = -\infty$, is shown at the bottom right of Figure 1.4.

Table 1.2 Table of the coefficients $\mathbf{w}^\star$ for $M = 9$ polynomials with various values for the regularization parameter λ. Note that $\ln\lambda = -\infty$ corresponds to a model with no regularization, i.e., to the graph at the bottom right in Figure 1.4. We see that, as the value of λ increases, the typical magnitude of the coefficients gets smaller.

	$\ln\lambda = -\infty$	$\ln\lambda = -18$	$\ln\lambda = 0$
$w_0^\star$	0.35	0.35	0.13
$w_1^\star$	232.37	4.74	-0.05
$w_2^\star$	-5321.83	-0.77	-0.06
$w_3^\star$	48568.31	-31.97	-0.05
$w_4^\star$	-231639.30	-3.89	-0.03
$w_5^\star$	640042.26	55.28	-0.02
$w_6^\star$	-1061800.52	41.32	-0.01
$w_7^\star$	1042400.18	-45.95	-0.00
$w_8^\star$	-557682.99	-91.53	0.00
$w_9^\star$	125201.43	72.68	0.01

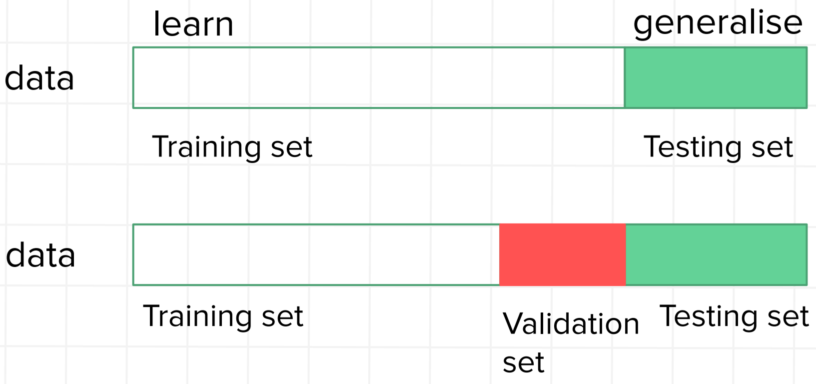
Graph of the root-mean-square error (1.3) versus $\ln\lambda$ for the $M = 9$ polynomial.



Model selection (an empirical view)

Minimizing square error / maximizing data likelihood can be a poor indication of performance on new data (generalisation) – Cause: overfitting

In the curve-fitting example: the order of the polynomial controls the number of free parameters in the model and thereby governs the model complexity.



How reliable are the estimates for validation and generalisation performance?

Figure 1.18 The technique of S-fold cross-validation, illustrated here for the case of $S = 4$, involves taking the available data and partitioning it into S groups (in the simplest case these are of equal size). Then $S - 1$ of the groups are used to train a set of models that are then evaluated on the remaining group. This procedure is then repeated for all S possible choices for the held-out group, indicated here by the red blocks, and the performance scores from the S runs are then averaged.

Figure 8.13 Nested cross-validation. We perform two levels of K-fold cross-validation.

[source: MML book]

What we did so far

ML 101:
Polynomial curve fitting: model, loss/error function, over-fitting, regularisation

Model selection

Probabilities: sum rule, product rule, Bayes theorem

Gaussians - 1D, maximum likelihood estimates (MLE), bias-variance
➡ and how this helps curve-fitting

Gaussians (multidimensional)
various matrix identities, geometric intuitions

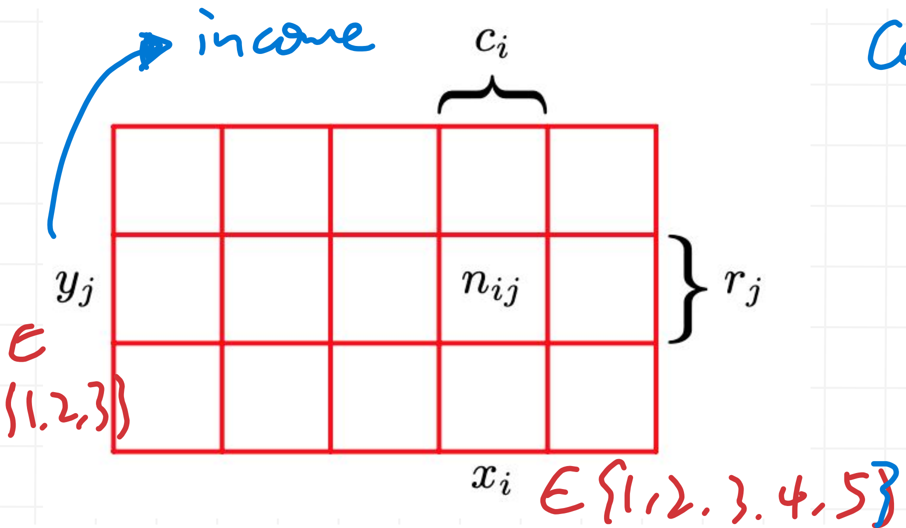
Bernoulli, Binomial, Exponential family distributions

Review: probabilities, derivatives and finding stationary points, eigen values and eigen vectors

The Rules of Probability

sum rule $$p(X) = \sum_{Y} p(X, Y) \tag{1.10}$$

product rule $$p(X, Y) = p(Y|X)p(X). \tag{1.11}$$

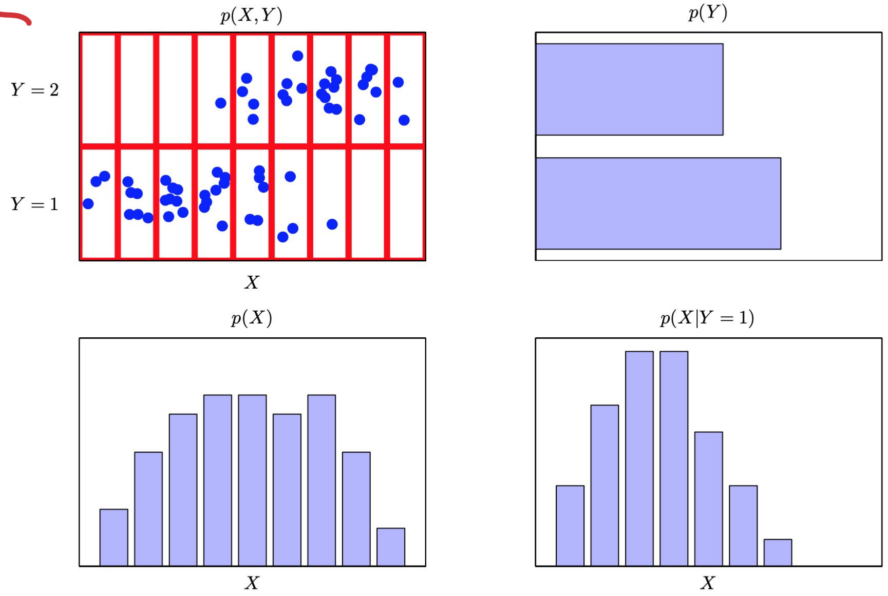


Count n_{ij} N sample.

$$P(X=x_i) = \frac{c_i}{N} = \frac{1}{N}\sum_{j=1}^{3} n_{ij}$$

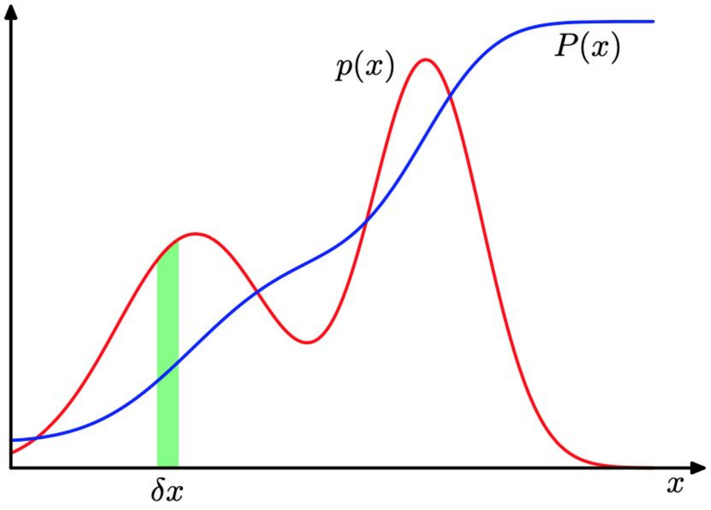
Bayes Theorem

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \tag{1.12}$$



Continuous random variables

Figure 1.12 The concept of probability for discrete variables can be extended to that of a probability density $p(x)$ over a continuous variable x and is such that the probability of x lying in the interval $(x, x + \delta x)$ is given by $p(x)\delta x$ for $\delta x \to 0$. The probability density can be expressed as the derivative of a cumulative distribution function $P(x)$.



$$p(x) = \int p(x, y)\,\mathrm{d}y \tag{1.31}$$

$$p(x, y) = p(y|x)p(x). \tag{1.32}$$

Bayes Theorem, restated (Sec 1.2.3)

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \quad (1.12)$$

likelihood. prior

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})p(\mathbf{w})}{p(\mathcal{D})} \quad (1.43)$$

not random once data is given

weight param data $\{(x_i, t_i),\ i=1,\cdots,N\}$

posterior ∝ likelihood × prior

Expectations, variance, covariance

For review

$$\mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x. \quad (1.34)$$

$x \sim p(x)$

$$\mathrm{var}[f] = \mathbb{E}\left[(f(x) - \mathbb{E}[f(x)])^2\right] \quad (1.38)$$

show this yourself

$x \sim p(x)$

$$= \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2. \quad (1.39)$$

$$\begin{aligned} \mathrm{cov}[x, y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y] \end{aligned} \quad (1.41)$$

$x, y \sim p(x,y)$

what is the expectation taken over? probability p is often implicit.

$Y=y$

$$\mathbb{E}_x[f|y] = \sum_x p(x|y) f(x) \quad (1.37)$$

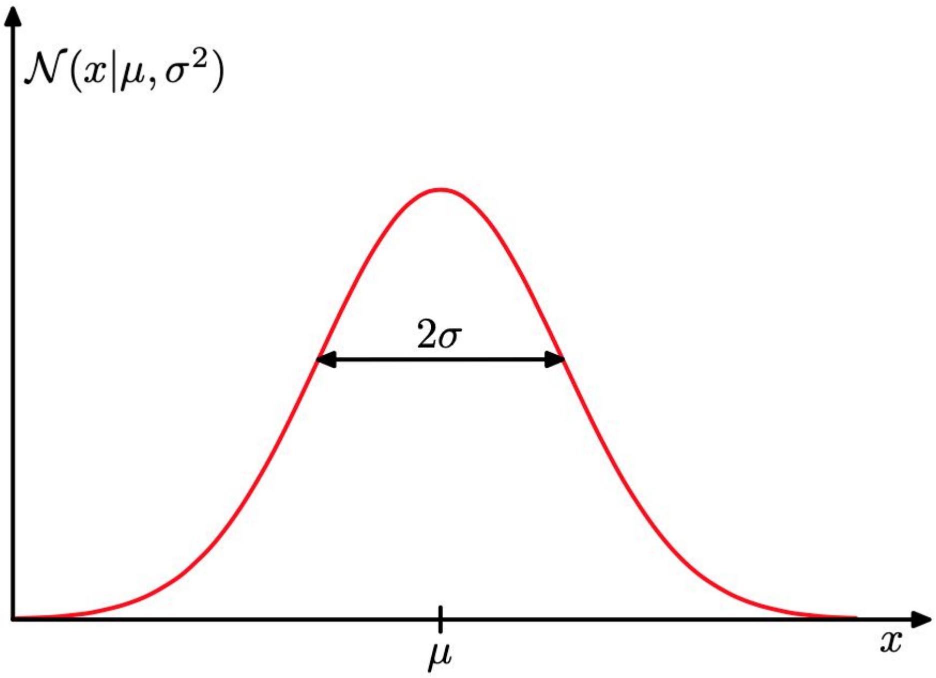
$f(x)$

$\rightarrow$ function of y

Question: for a random variable x ~ p(x), do E[x] and var[x] always exist? No.

The Gaussian Distribution

Figure 1.13 Plot of the univariate Gaussian showing the mean μ and the standard deviation σ.



$$\mathcal{N}\left(x|\mu,\sigma^2\right) = \underbrace{\frac{1}{(2\pi\sigma^2)^{1/2}}}_{\text{normalising}} \exp\left\{ \underbrace{-\frac{1}{2\sigma^2}(x-\mu)^2}_{\text{quadratic term}} \right\}$$

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu}) \right\}$$

Maximum likelihood for univariate Gaussian

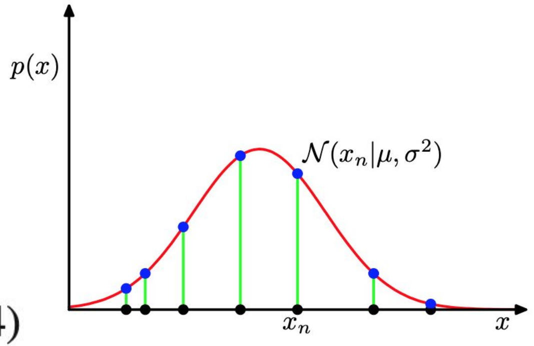
$\max_{\mu,\sigma^2}$

$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\underbrace{\mu,\sigma^2}_{}\right).$$ data likelihood

variable.

$\times \to +$

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \cancel{\frac{N}{2}\ln\sigma^2} - \cancel{\frac{N}{2}\ln(2\pi)}. \qquad (1.54)$$



$$\mu_{\mathrm{ML}} = \quad \frac{\partial L}{\partial \mu} = \cancel{-\frac{1}{2\sigma^2}}\cdot 2\sum_{n=1}^{N}(x_n-\mu)\cancel{(-1)} = 0 \Rightarrow \sum_{n=1}^{N} x_n - N\mu = 0$$

$$\sigma^2_{\mathrm{ML}} = \quad \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Maximum likelihood $\neq$ *unbiased*

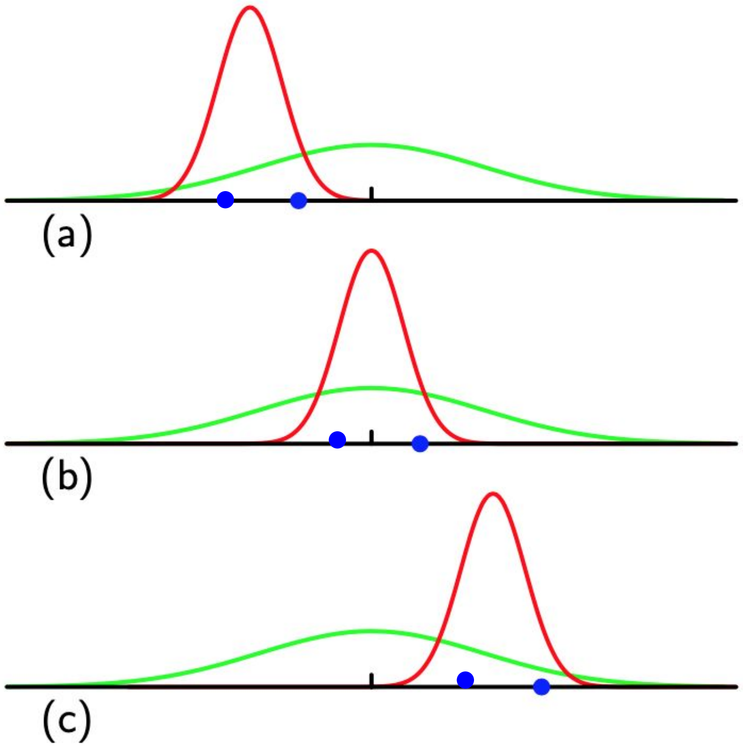
In statistics, the **bias** (or **bias function**) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated. An estimator or decision rule with zero bias is called **unbiased**. In statistics, "bias" is an **objective** property of an estimator.

"Bias" is not necessarily bad!

$$\boxed{\begin{aligned} \mathbb{E}[\mu_{\mathrm{ML}}] &= \mu \qquad (1.57) \\ \mathbb{E}[\sigma^2_{\mathrm{ML}}] &= \left(\frac{N-1}{N}\right)\sigma^2 \qquad (1.58) \end{aligned}}$$

$$\boxed{\widetilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \mu_{\mathrm{ML}})^2.} \qquad (1.59)$$

Figure 1.15 Illustration of how bias arises in using maximum likelihood to determine the variance of a Gaussian. The green curve shows the true Gaussian distribution from which data is generated, and the three red curves show the Gaussian distributions obtained by fitting to three data sets, each consisting of two data points shown in blue, using the maximum likelihood results (1.55) and (1.56). Averaged across the three data sets, the mean is correct, but the variance is systematically under-estimated because it is measured relative to the sample mean and not relative to the true mean.

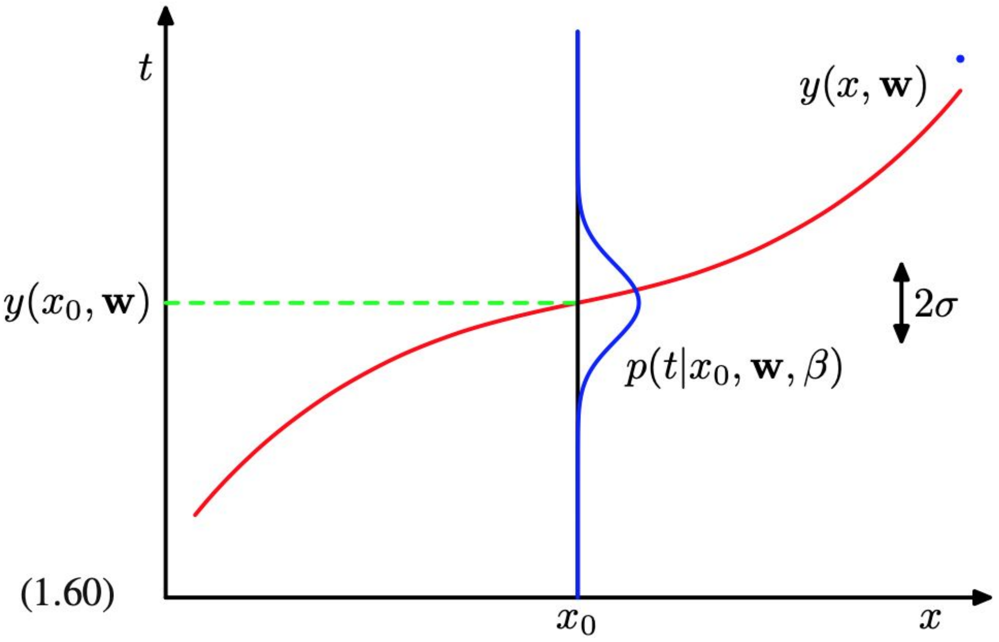


Q: does high bias/variance means that the model is overfitted, or vice versa?

Bringing it together:

Curve fitting with maximum likelihood

Figure 1.16 Schematic illustration of a Gaussian conditional distribution for t given x given by (1.60), in which the mean is given by the polynomial function $y(x, \mathbf{w})$, and the precision is given by the parameter β, which is related to the variance by $\beta^{-1} = \sigma^2$.



$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \qquad (1.60)$$

Goal: estimate \beta β

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}\left(t|y(x, \mathbf{w}), \beta^{-1}\right) \tag{1.60}$$

Gaussian.

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}\left(t_n|y(x_n, \mathbf{w}), \beta^{-1}\right). \tag{1.61}$$

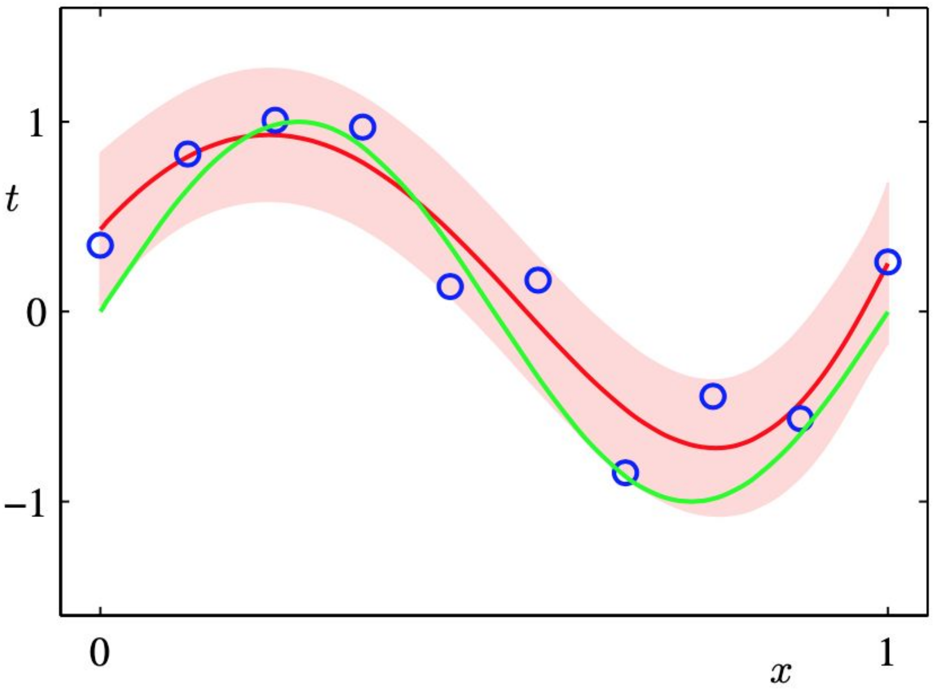
$$\ln p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = -\frac{\beta}{2} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}) - t_n\}^2 + \frac{N}{2} \ln \beta - \frac{N}{2} \ln(2\pi). \tag{1.62}$$

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{1}{N} \sum_{n=1}^{N} \{y(x_n, \mathbf{w}_{\mathrm{ML}}) - t_n\}^2. \tag{1.63}$$

curve-fitting: predictive distribution

(will cover next week in Bayesian linear regression)

Figure 1.17 The predictive distribution resulting from a Bayesian treatment of polynomial curve fitting using an $M = 9$ polynomial, with the fixed parameters $\alpha = 5 \times 10^{-3}$ and $\beta = 11.1$ (corresponding to the known noise variance), in which the red curve denotes the mean of the predictive distribution and the red region corresponds to ± 1 standard deviation around the mean.



What we did so far

ML 101:
Polynomial curve fitting: model, loss/error function, over-fitting, regularisation

Probabilities: sum rule, product rule, Bayes theorem
Gaussians - 1D, MLE, bias-variance
➜ and how this helps curve-fitting

Bernoulli, binomial

Gaussians (multidimensional)
various matrix identities, geometric intuitions

Exponential family

Review: probabilities, derivatives and finding stationary points, eigen values and eigen vectors

From Bernoulli to Binomial

$$\mathrm{Bern}(x|\mu) = \mu^x(1-\mu)^{1-x}$$

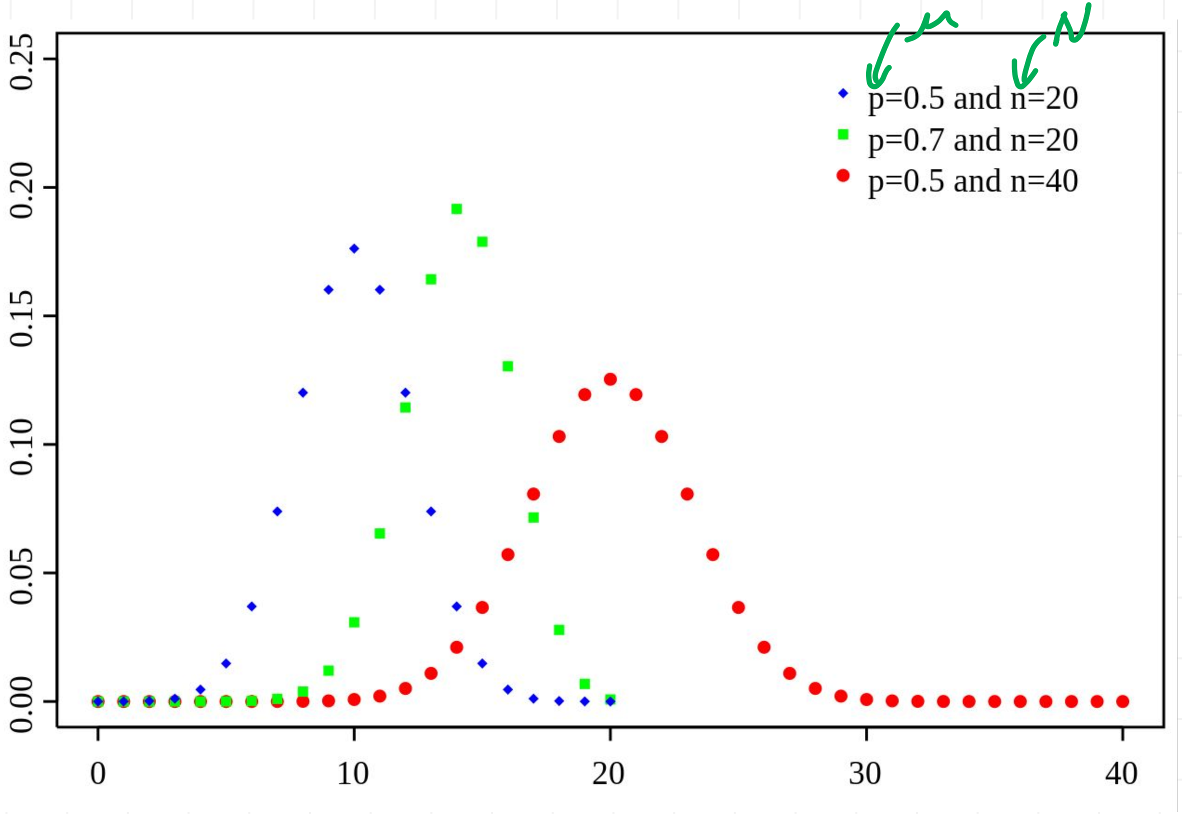
$x \in \{0,1\}$

N tosses

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m}\mu^m(1-\mu)^{N-m} \tag{2.9}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{2.10}$$

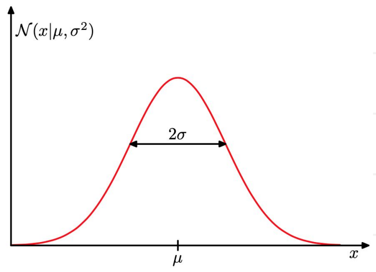
binomial for increasing large N



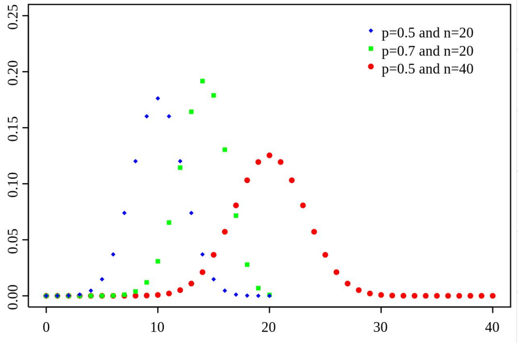
Gaussians again: why is it cool?

CLT - central limit theorem

max entropy



n coin tosses with p



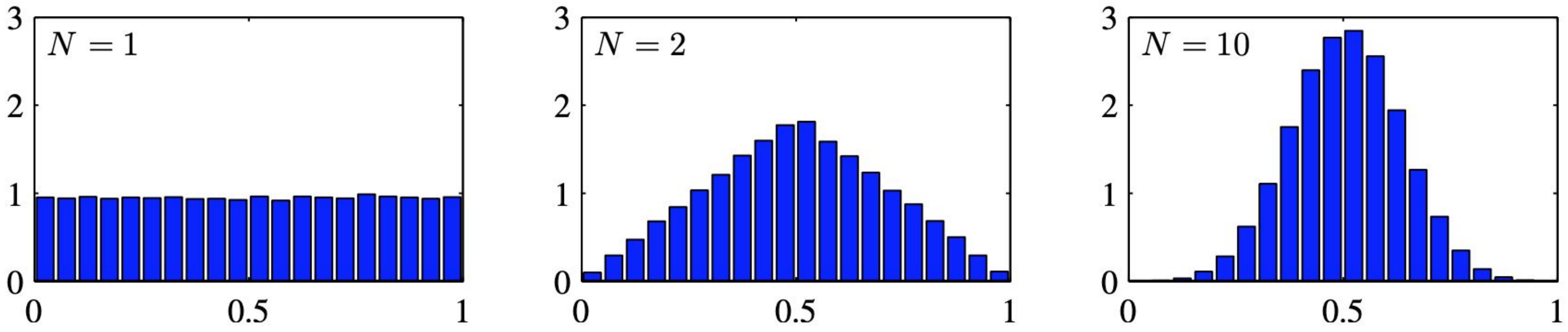
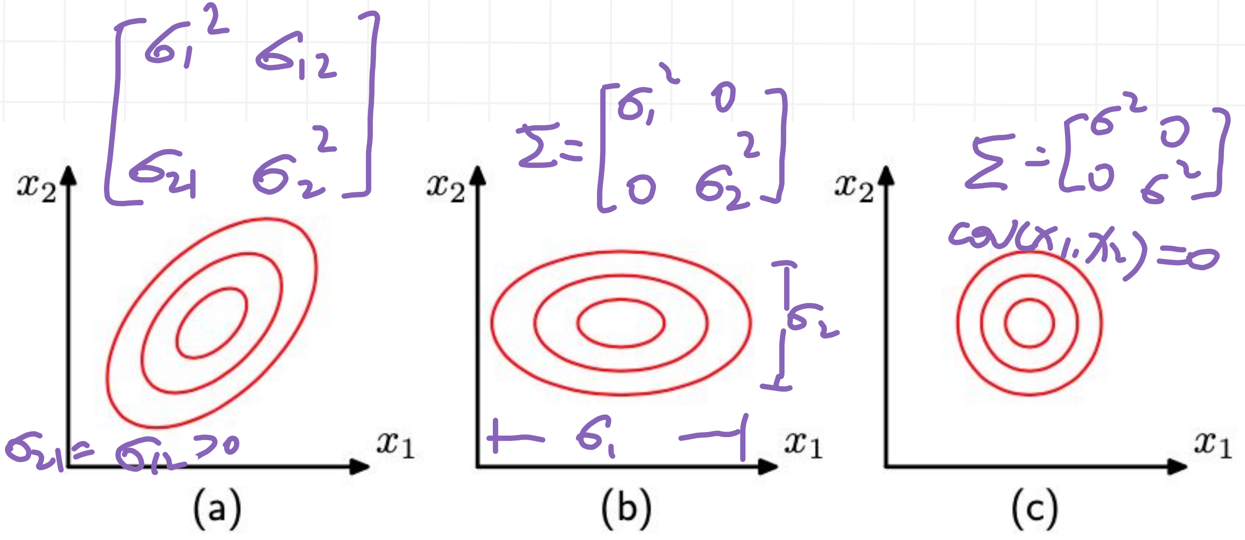


Figure 2.6 Histogram plots of the mean of N uniformly distributed numbers for various values of N. We observe that as N increases, the distribution tends towards a Gaussian.

Gaussians – multidimensional

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \quad (2.43)$$

Figure 2.8 Contours of constant probability density for a Gaussian distribution in two dimensions in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to the identity matrix, in which the contours are concentric circles.



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{\underbrace{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}_{\text{squared distance}}\right\} \tag{2.43}$$

The Mahalanobis distance

Eigen decomposition of the cov matrix

$\lambda_i \geq 0$

$$\boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i\mathbf{u}_i \tag{2.45}$$

orthonormal.

$u_i^T u_i = 1$

$u_i^T u_j = 0, \ i \neq j$

$$\mathbf{u}_i^{\mathrm{T}}\mathbf{u}_j = I_{ij} \tag{2.46}$$

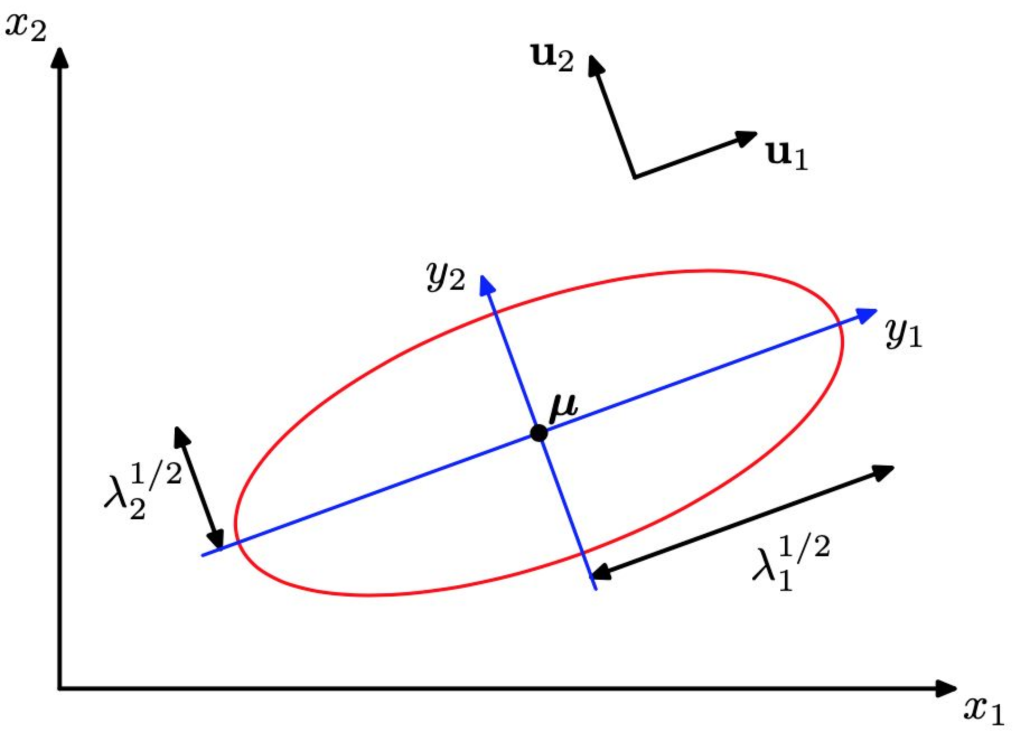
$$\boldsymbol{\Sigma} = \sum_{i=1}^{D}\lambda_i\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}} \qquad \boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D}\frac{1}{\lambda_i}\mathbf{u}_i\mathbf{u}_i^{\mathrm{T}}.$$

$$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = \sum_{i=1}^{D}\frac{y_i^2}{\lambda_i} \qquad y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$$

Contours of general 2-D Gaussians - a rotated ellipse

$$\mathbf{y} = \mathbf{U}(\mathbf{x} - \boldsymbol{\mu})$$

Figure 2.7 The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



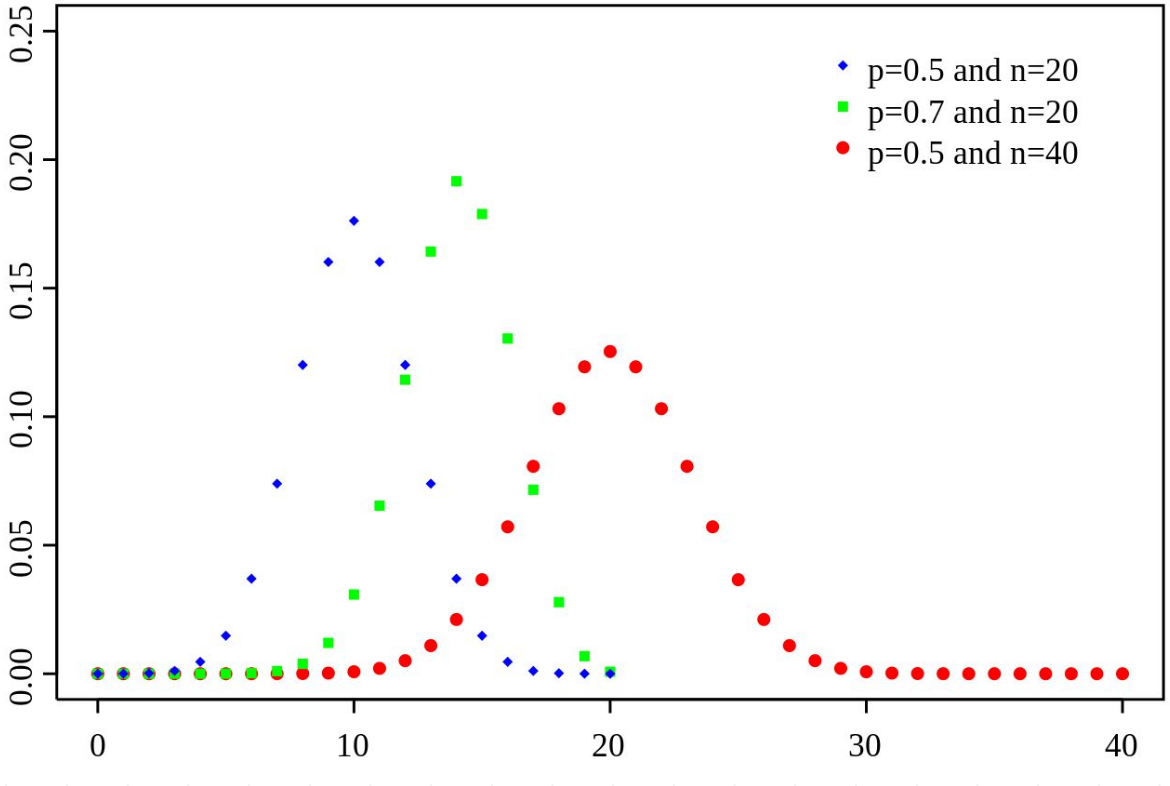
From Bernoulli to Binomial

$$\mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x}$$

$$\mathrm{Bin}(m|N,\mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m} \tag{2.9}$$

$$\binom{N}{m} \equiv \frac{N!}{(N-m)!m!} \tag{2.10}$$

binomial for increasing large N



The Exponential family

Beyond Gaussians: What is a class of ‘nice’ distributions for statistical machine learning?

- More expressive
- “Easy” to estimate

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \qquad (2.194)$$

normalisation

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \mathbf{u}(\mathbf{x})\right\} \mathrm{d}\mathbf{x} = 1 \tag{2.195}$$

Special case: Bernoulli

$$p(x|\mu) = \mathrm{Bern}(x|\mu) = \mu^{x}(1-\mu)^{1-x}. \tag{2.196}$$

$$\begin{aligned} &= \exp\left\{x \ln \mu + (1-x)\ln(1-\mu)\right\} \\ &= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}. \end{aligned} \tag{2.197}$$

$$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$$

$$p(x|\eta) = \sigma(-\eta)\exp(\eta x) \tag{2.200}$$

Special case: Gaussian

$$p(x|\mu,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{2.218}$$

$$= \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}x^2 + \frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}\mu^2\right\} \tag{2.219}$$

$$\boldsymbol{\eta} = \begin{pmatrix} \mu/\sigma^2 \\ -1/2\sigma^2 \end{pmatrix} \tag{2.220}$$

$$\mathbf{u}(x) = \begin{pmatrix} x \\ x^2 \end{pmatrix} \tag{2.221}$$

$$h(\mathbf{x}) = (2\pi)^{-1/2} \tag{2.222}$$

$$g(\boldsymbol{\eta}) = (-2\eta_2)^{1/2} \exp\left(\frac{\eta_1^2}{4\eta_2}\right). \tag{2.223}$$

normalisation

$$g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \,\mathrm{d}\mathbf{x} = 1 \tag{2.195}$$

$$\nabla g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \,\mathrm{d}\mathbf{x} + g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \,\mathrm{d}\mathbf{x} = 0. \tag{2.224}$$

$$-\frac{1}{g(\boldsymbol{\eta})}\nabla g(\boldsymbol{\eta}) = g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x}) \,\mathrm{d}\mathbf{x} = \mathbb{E}[\mathbf{u}(\mathbf{x})] \tag{2.225}$$

MLE and sufficient stats

$$-\nabla \ln g(\boldsymbol{\eta}_{\mathrm{ML}}) = \frac{1}{N}\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

Exponential family: a note about notations

$$p(\mathbf{x}|\boldsymbol{\eta}) = h(\mathbf{x})g(\boldsymbol{\eta}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \tag{2.194}$$

h(x)=1, g(η) = exp(-ψ(**η**))

Assignment 1 Q2

Definition 1 (Exponential Family[2]). Given a function $\boldsymbol{u} : \mathbb{R} \to \mathbb{R}^m$, we denote an exponential family distribution as $\textsc{Exp}(\boldsymbol{u}, \boldsymbol{\eta})$, where $\boldsymbol{\eta} \in \mathcal{P} \subset \mathbb{R}^m$ designates the m-dimensional parameters of the distribution within an exponential family[3]. The corresponding densities of the distributions are given by

$$q(x; \boldsymbol{\eta}) = \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x) - \psi(\boldsymbol{\eta})), \tag{1}$$

where

$$\psi(\boldsymbol{\eta}) = \log \int \exp(\boldsymbol{\eta}^\top \boldsymbol{u}(x)) \, \mathrm{d}x. \tag{2}$$

The function $\boldsymbol{u}$ is called the sufficient statistics of the exponential family and the function ψ is called the log-partition function of the exponential family.

About these lecture notes:

- They are designed to be visual aid but not reading material (you have the book for that).
- They are generally focused on derivations + plots and less on the “story” part of the model.
- I do not aim to produce new equations nor new plots (they don’t necessarily help you learn :)

A word about data/plots in the book:

- Reasoning about ML models on toy data is a core skill of a good ML engineer.
- Designing appropriate toy data is a core research skill in ML.

What we covered today

ML 101:
Polynomial curve fitting: model, loss/error function, over-fitting, regularisation

Model selection

Probabilities: sum rule, product rule, Bayes theorem
Gaussians - 1D, MLE, bias-variance
➡ and how this helps curve-fitting

Gaussians (multidimensional)
various matrix identities, geometric intuitions

Bernoulli, Binomial, Exponential family distributions

Review: probabilities, derivatives and finding stationary points, eigen values and eigen vectors