

I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER  
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE  
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

# Announcements

Mon 14 Mar week 4 - no lecture, Canberra day!

Assignment 1, submit pdf + code – gradescope entry will be up.

More about this 1630-1700 today

- You're encouraged to typeset the solution with Latex, since making your computer "speak math" (and manage bibliography) is one of the important skills for AI/ML. We liken this to using source control in COMP1100/1110.  
→ 2 bonus marks for solutions made with Latex
- Grading expectations for COMP4670 students –  $\max\{\text{COMP4670 scheme, COMP8600 scheme}\}$

# Linear models for regression 2

We covered:

- Basis functions
- Maximum Likelihood with Gaussian Noise
- Regularisation
- Bias variance decomposition

Mon

$$\mathbf{w} = (\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}. \quad (3.28)$$

Today:

Conditional gaussians

We'll get there via 3 stepping stones - conjugate prior, conditional Gaussian, Bayes theorem for Gaussian variables

Bayesian regression

Predictive distributions

Curse of dimensionality

Info theory 101

### Definition ( Conjugate Prior)

A class of prior probability distributions  $p(w)$  is conjugate to a class of likelihood functions  $p(x|w)$  if the resulting posterior distributions  $p(w|x)$  are in the same family as  $p(w)$ .

Table: Discrete likelihood distributions

Likelihood	Conjugate Prior
Bernoulli	Beta
Binomial	Beta
Poisson	Gamma
Multinomial	Dirichlet

Table: Continuous likelihood distributions

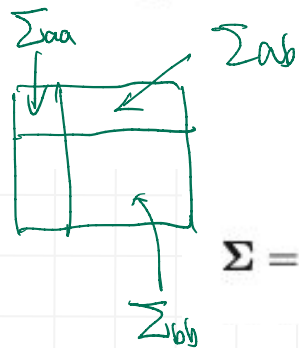
Likelihood	Conjugate Prior
Uniform	Pareto
Exponential	Gamma
Normal	Normal (mean parameter)
Multivariate normal	Multivariate normal (mean parameter)

$p(w)$   
 $p(w|x)$

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

## Partitioned Gaussians

Given a joint Gaussian distribution  $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$  with  $\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1}$  and



$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \quad (2.94)$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}, \quad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}. \quad (2.95)$$

Marginal distribution:

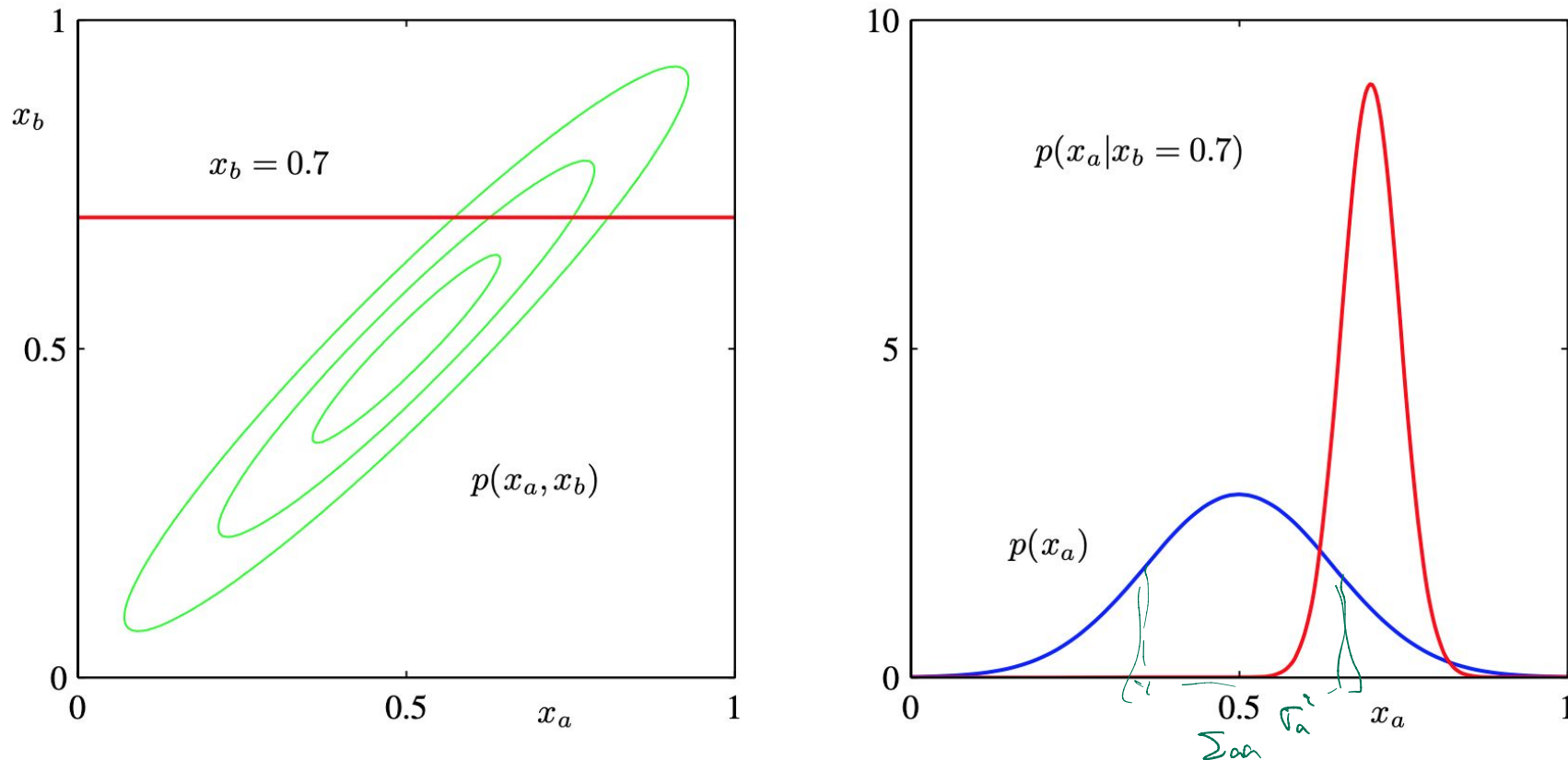
$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b$$

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a | \boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa}). \quad (2.98)$$

What about the conditional distribution?

$$p(\mathbf{x}_a | \mathbf{x}_b)$$

## Intuitions for the conditional and marginal of jointly Gaussian variables



**Figure 2.9** The plot on the left shows the contours of a Gaussian distribution  $p(x_a, x_b)$  over two variables, and the plot on the right shows the marginal distribution  $p(x_a)$  (blue curve) and the conditional distribution  $p(x_a | x_b)$  for  $x_b = 0.7$  (red curve).

# Covariance and precision matrices

$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \quad (2.78)$$

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \\ -\mathbf{D}^{-1}\mathbf{C}\mathbf{M} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{C}\mathbf{M}\mathbf{B}\mathbf{D}^{-1} \end{pmatrix} \quad (2.76)$$

where we have defined

$$\mathbf{M} = (\mathbf{A} - \mathbf{B}\mathbf{D}^{-1}\mathbf{C})^{-1}. \quad (2.77)$$

The quantity  $\mathbf{M}^{-1}$  is known as the *Schur complement* of the matrix on the left-hand side of (2.76) with respect to the submatrix  $\mathbf{D}$ .

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \quad (2.79)$$

$$\Lambda_{ab} = -(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1}\Sigma_{ab}\Sigma_{bb}^{-1}. \quad (2.80)$$

# “Completing the square”

A useful view for manipulating Gaussian distributions

$\log ( N (x \mid \mu, \Sigma) ) = \text{const} + \log ( \exp ( \text{quadratic expression of } x ) )$

$$-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) = -\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{-1}\mathbf{x} + \mathbf{x}^T \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} + \text{const} \quad (2.71)$$



$$\begin{pmatrix} \Sigma_{aa} & \Sigma_{ab} \\ \Sigma_{ba} & \Sigma_{bb} \end{pmatrix}^{-1} = \begin{pmatrix} \Lambda_{aa} & \Lambda_{ab} \\ \Lambda_{ba} & \Lambda_{bb} \end{pmatrix} \quad (2.78)$$

“Completing the square”

$$\Lambda_{aa} = (\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})^{-1} \quad (2.79)$$

$$\begin{aligned} -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}) &= \text{quadratic in } x_a \\ &= -\frac{1}{2}(\underline{x_a} - \underline{\mu_a})^T \underline{\Lambda_{aa}}(\underline{x_a} - \underline{\mu_a}) - \frac{1}{2}(\underline{x_a} - \underline{\mu_a})^T \underline{\Lambda_{ab}}(\underline{x_b} - \underline{\mu_b}) \\ &\quad - \frac{1}{2}(\underline{x_b} - \underline{\mu_b})^T \underline{\Lambda_{ba}}(\underline{x_a} - \underline{\mu_a}) - \frac{1}{2}(\underline{x_b} - \underline{\mu_b})^T \underline{\Lambda_{bb}}(\underline{x_b} - \underline{\mu_b}). \end{aligned} \quad (2.70)$$

linear in  $x_a$   
const.

Consider

$$p(\mathbf{x}_a | \mathbf{x}_b)$$

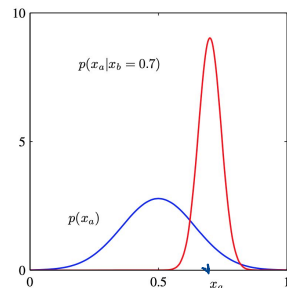
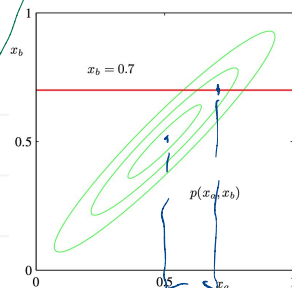
$$\Lambda_{ab} = \Lambda_{ba} \quad \text{const. } x_b$$

From the quadratic term:

$$\Sigma_{a|b} = \Lambda_{aa}^{-1}.$$

From the linear term:

$$\begin{aligned} \mu_{a|b} &= \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (x_b - \mu_b) \} \\ &= \mu_a - \underbrace{\Lambda_{aa}^{-1} \Lambda_{ab}}_{\text{shift on mean}} (x_b - \mu_b) \end{aligned} \quad \text{how far } x_b \text{ is from its mean} \quad (2.75)$$



(2.73) shift

Conditional distribution:

$$(\Sigma_{aa} - \Sigma_{ab}\Sigma_{bb}^{-1}\Sigma_{ba})$$

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Lambda}_{aa}^{-1}) \quad (2.96)$$

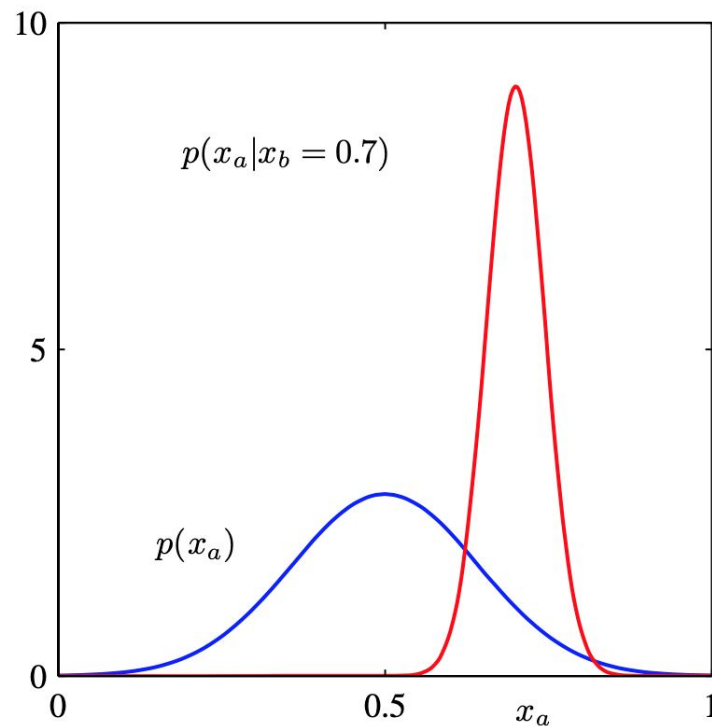
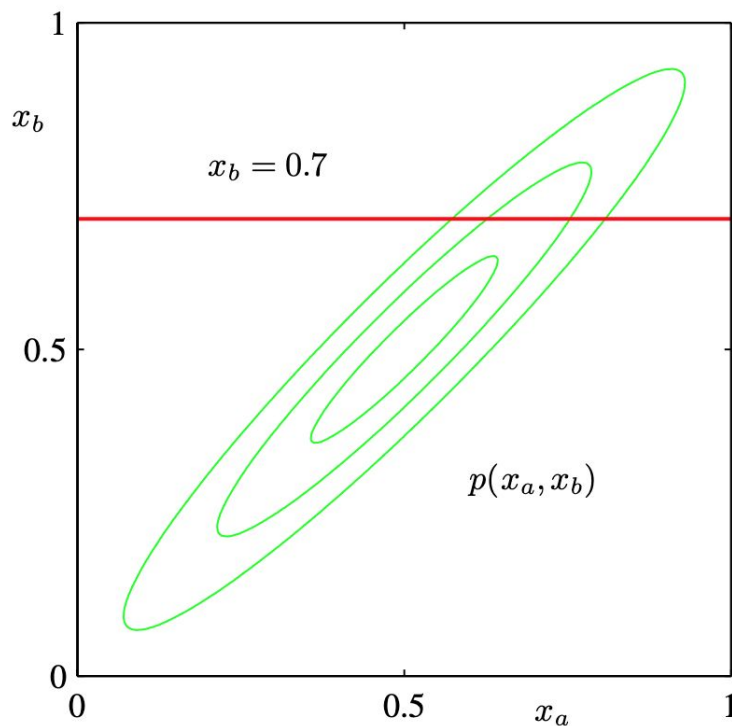
$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b). \quad (2.97)$$

Marginal distribution:

$$p(\mathbf{x}_a) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \Sigma_{aa}). \quad (2.98)$$

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_{a|b}, (\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}))$$

$$\boldsymbol{\mu}_{a|b} = \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b).$$



**Figure 2.9** The plot on the left shows the contours of a Gaussian distribution  $p(x_a, x_b)$  over two variables, and the plot on the right shows the marginal distribution  $p(x_a)$  (blue curve) and the conditional distribution  $p(x_a|x_b)$  for  $x_b = 0.7$  (red curve).

# Bayes theorem for Gaussian variables

Given a marginal Gaussian distribution for  $\mathbf{x}$  and a conditional Gaussian distribution for  $\mathbf{y}$  given  $\mathbf{x}$  in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad p(\mathbf{t} | \mathbf{y}(\mathbf{x}, \mathbf{u}))^{\sigma^{-1}} \quad (2.113)$$

$$p(\mathbf{y} | \mathbf{x}) = \mathcal{N}(\mathbf{y} | \mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of  $\mathbf{y}$  and the conditional distribution of  $\mathbf{x}$  given  $\mathbf{y}$  are given by

Hunch:

$$\boldsymbol{\mu}_y = \mathbf{A}\boldsymbol{\mu} + \mathbf{b}$$

$$\sigma_y^2 : \text{involve } \boldsymbol{\Lambda}^{-1}, \mathbf{L}^{-1}$$

$$p(\mathbf{y}) =$$

$$p(\mathbf{x} | \mathbf{y}) =$$

$$\begin{matrix} \mu_{x|y} & \mu_{A,b} \\ \sigma_y^2 & \boldsymbol{\Lambda}^{-1} \mathbf{L}^{-1} \end{matrix}$$

General direction: find the joint of  $p(\mathbf{x}, \mathbf{y})$ , and then use results from conditional+marginal Gaussians.

$$\mathbf{z} = \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}$$

$$\begin{aligned} p(\mathbf{x}) &= \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \\ p(\mathbf{y}|\mathbf{x}) &= \mathcal{N}(\mathbf{y}|\mathbf{Ax} + \mathbf{b}, \mathbf{L}^{-1}) \end{aligned}$$

$$\begin{aligned} \ln p(\mathbf{z}) &= \ln p(\mathbf{x}) + \ln p(\mathbf{y}|\mathbf{x}) \\ &= -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Lambda}(\mathbf{x} - \boldsymbol{\mu}) \\ &\quad -\frac{1}{2}(\mathbf{y} - \mathbf{Ax} - \mathbf{b})^T \mathbf{L}(\mathbf{y} - \mathbf{Ax} - \mathbf{b}) + \underbrace{\text{const}} \end{aligned} \quad (2.102)$$

$$\begin{aligned} &-\frac{1}{2}\mathbf{x}^T(\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{LA})\mathbf{x} - \frac{1}{2}\mathbf{y}^T\mathbf{Ly} + \frac{1}{2}\mathbf{y}^T\mathbf{LAx} + \frac{1}{2}\mathbf{x}^T\mathbf{A}^T\mathbf{Ly} \\ &= -\frac{1}{2}\begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}^T \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{LA} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{LA} & \mathbf{L} \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = -\frac{1}{2}\mathbf{z}^T\mathbf{Rz} \end{aligned} \quad (2.103)$$

$$\mathbf{R} = \begin{pmatrix} \boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{LA} & -\mathbf{A}^T\mathbf{L} \\ -\mathbf{LA} & \mathbf{L} \end{pmatrix}. \quad (2.104)$$

$$\mathbf{R} = \begin{pmatrix} \mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A} & -\mathbf{A}^T \mathbf{L} \\ -\mathbf{L} \mathbf{A} & \mathbf{L} \end{pmatrix}. \quad (2.104)$$

$$\mathbb{E}[\mathbf{z}] = \mathbf{R}^{-1} \begin{pmatrix} \mathbf{\Lambda} \boldsymbol{\mu} - \mathbf{A}^T \mathbf{L} \mathbf{b} \\ \mathbf{L} \mathbf{b} \end{pmatrix}. \quad \mathbb{E}[\mathbf{z}] = \begin{pmatrix} \boldsymbol{\mu} \\ \mathbf{A} \boldsymbol{\mu} + \mathbf{b} \end{pmatrix}. \quad (2.108)$$

recall

$$\Sigma_{a|b} = \Lambda_{aa}^{-1}. \quad (2.73)$$

$$\begin{aligned} \mu_{a|b} &= \Sigma_{a|b} \{ \Lambda_{aa} \mu_a - \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \} \\ &= \mu_a - \Lambda_{aa}^{-1} \Lambda_{ab} (\mathbf{x}_b - \boldsymbol{\mu}_b) \end{aligned} \quad (2.75)$$

$$\mathbb{E}[\mathbf{x}|\mathbf{y}] = (\mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1} \{ \mathbf{A}^T \mathbf{L} (\mathbf{y} - \mathbf{b}) + \mathbf{\Lambda} \boldsymbol{\mu} \} \quad (2.111)$$

$$\text{cov}[\mathbf{x}|\mathbf{y}] = (\mathbf{\Lambda} + \mathbf{A}^T \mathbf{L} \mathbf{A})^{-1}. \quad (2.112)$$

# Bayes theorem for Gaussian variables

## Marginal and Conditional Gaussians

Given a marginal Gaussian distribution for  $\mathbf{x}$  and a conditional Gaussian distribution for  $\mathbf{y}$  given  $\mathbf{x}$  in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of  $\mathbf{y}$  and the conditional distribution of  $\mathbf{x}$  given  $\mathbf{y}$  are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

# What we have covered

Prelude: Conjugate prior, conditional of Gaussian variables, Bayes Theorem of Gaussian distributions

Bayesian regression

Predictive distributions

Curse of dimensionality

Info theory 101



Goal: estimate the posterior of  $\mathbf{w}$ , not just  $\mathbf{w}_{\text{ML}}$

if

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \phi(\mathbf{x}_n), \beta^{-1}) \quad (3.10)$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0) \quad (3.48)$$

then

posterior after seeing  $N$  data points  $\{(x_i, t_i), i=1, \dots, N\}$

$$p(\mathbf{w} | \mathbf{t}) = \mathcal{N}(\mathbf{w} | \mathbf{m}_N, \mathbf{S}_N) \quad (3.49)$$

$$\mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) \quad (3.50)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \Phi^T \Phi. \quad (3.51)$$

## Marginal and Conditional Gaussians

for linear reg,

Given a marginal Gaussian distribution for  $\mathbf{x}$  and a conditional Gaussian distribution for  $\mathbf{y}$  given  $\mathbf{x}$  in the form

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Lambda}^{-1}) \quad (2.113)$$

$$p(\mathbf{y}|\mathbf{x}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\mathbf{x} + \mathbf{b}, \mathbf{L}^{-1}) \quad (2.114)$$

the marginal distribution of  $\mathbf{y}$  and the conditional distribution of  $\mathbf{x}$  given  $\mathbf{y}$  are given by

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.115)$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Sigma}\{\mathbf{A}^T\mathbf{L}(\mathbf{y} - \mathbf{b}) + \boldsymbol{\Lambda}\boldsymbol{\mu}\}, \boldsymbol{\Sigma}) \quad (2.116)$$

where

$$\boldsymbol{\Sigma} = (\boldsymbol{\Lambda} + \mathbf{A}^T\mathbf{L}\mathbf{A})^{-1}. \quad (2.117)$$

assume

$$p(\mathbf{t}|\mathbf{X}, \mathbf{w}, \beta) = \prod_{n=1}^N \mathcal{N}(t_n|\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \quad (3.10)$$

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_0, \mathbf{S}_0) \quad (3.48)$$

then

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \quad (3.49)$$

$$\mathbf{m}_N = \mathbf{S}_N (\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t}) \quad (3.50)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}. \quad (3.51)$$

$$\mathbf{x} \rightarrow \mathbf{w}$$

$$\mathbf{y} \rightarrow \mathbf{t}$$

$$\mathbf{A} \rightarrow \boldsymbol{\Phi}_{N \times m}$$

# Isotropic Gaussian prior

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \quad (3.49)$$

general prior  
 $p(\mathbf{w}) \sim \mathcal{N}(\mathbf{m}_0, \mathbf{S}_0)$

$$\mathbf{m}_N = \frac{\mathbf{S}_N}{\mathbf{S}_0} (\mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \Phi^T \mathbf{t}) \quad (3.50)$$

$$\mathbf{S}_N^{-1} = \mathbf{S}_0^{-1} + \beta \Phi^T \Phi. \quad (3.51)$$

← "data covariance"  
 ← fixed, doesn't change after seeing data.

← prior ↑ if N small  
 data ↑ if N large.

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1} \mathbf{I}) \quad (3.52)$$

$$\mathbf{m}_N = \beta \mathbf{S}_N \Phi^T \mathbf{t} \quad (3.53)$$

$$\mathbf{S}_N^{-1} = (\alpha \mathbf{I} + \beta \Phi^T \Phi)^{-1} \quad (3.54)$$

$$\mathbf{w}_{\text{ML}} = (\Phi^T \Phi)^{-1} \Phi^T \mathbf{t} \quad (3.15)$$

# Bayesian Regression

- Log of posterior is sum of log likelihood and log of prior

$$\ln p(\mathbf{w} | \mathbf{t}) = -\beta \underbrace{\frac{1}{2} \|\mathbf{t} - \Phi \mathbf{w}\|^2}_{\text{sum-of-squares-error}} - \frac{\alpha}{2} \underbrace{\|\mathbf{w}\|^2}_{\text{regulariser}} + \text{const.}$$

- The *maximum a posteriori* estimator

$$\mathbf{w}_{\text{m.a.p.}} = \arg \max_{\mathbf{w}} \underline{p(\mathbf{w} | \mathbf{t})}$$

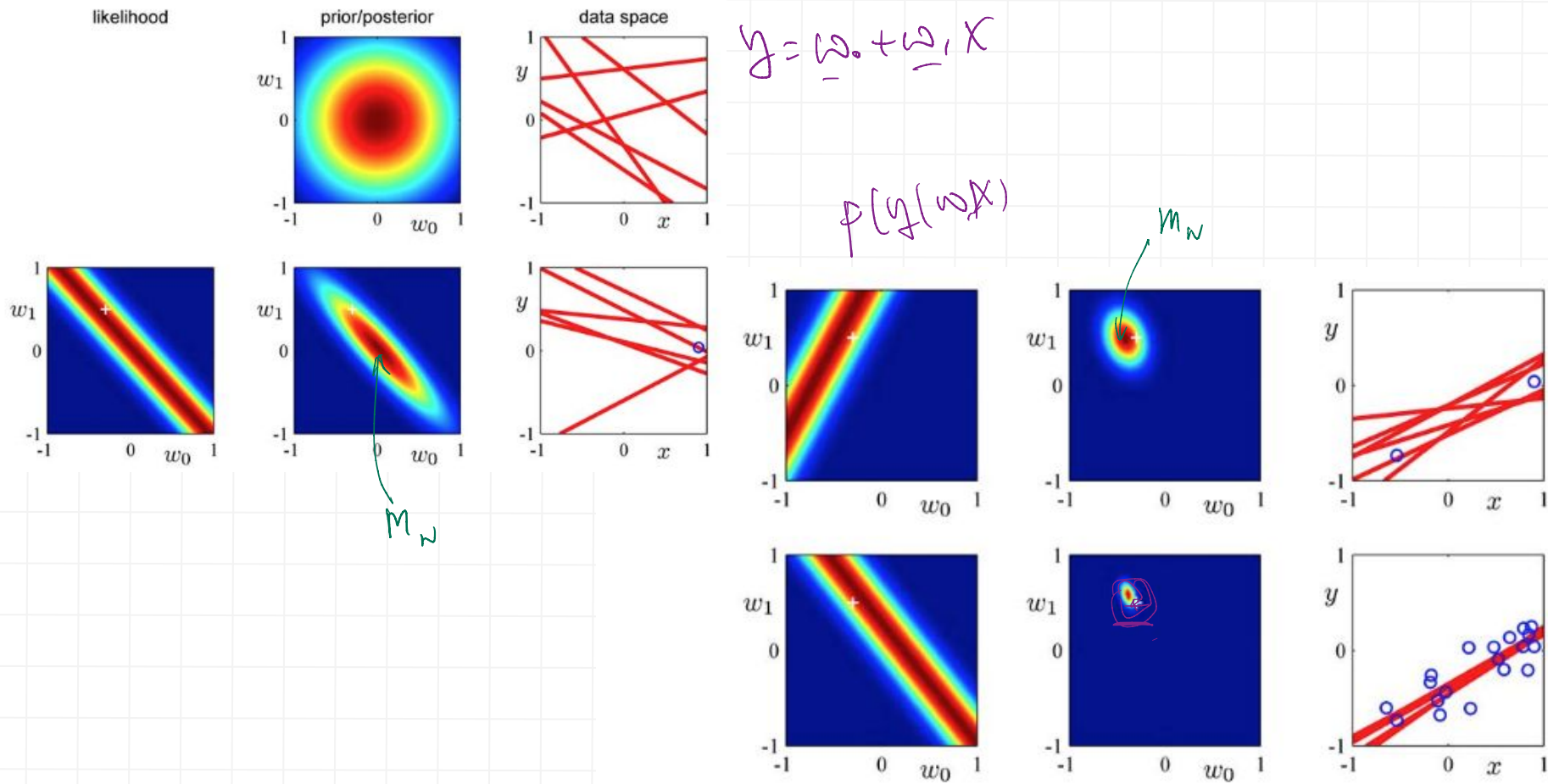
corresponds to minimising the sum-of-squares error function with quadratic regularisation coefficient  $\lambda = \alpha/\beta$ .

- The posterior is Gaussian so mode = mean:  $\mathbf{w}_{\text{m.a.p.}} = \mathbf{m}_N$ .
- For  $\alpha \ll \beta$  we recover unregularised least squares (equivalently m.a.p. approaches maximum likelihood), for example in case of
  - an infinitely broad prior with  $\alpha \rightarrow 0$
  - an infinitely precise likelihood with  $\beta \rightarrow \infty$

$$\begin{aligned} \mathbf{m}_N &= \underline{\beta \mathbf{S}_N \Phi^T \mathbf{t}} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \Phi^T \Phi. \end{aligned}$$



$$\mathbf{w} = \underline{(\lambda \mathbf{I} + \Phi^T \Phi)^{-1} \Phi^T \mathbf{t}.}$$



**Figure 3.7** Illustration of sequential Bayesian learning for a simple linear model of the form  $y(x, \mathbf{w}) = w_0 + w_1 x$ . A detailed description of this figure is given in the text.

# What we have covered

Prelude: Conjugate prior, conditional of Gaussian variables, Bayes Theorem of Gaussian distributions

Bayesian regression

Predictive distributions

Curse of dimensionality

Info theory 101

# Predictive distribution

Goal of regression: estimate  $y(x; \mathbf{w})$  at unobserved values of  $x$

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\mathbf{w}, \beta) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) d\mathbf{w} \quad (3.57)$$

$$p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(t|y(\mathbf{x}, \mathbf{w}), \beta^{-1}). \quad (3.8)$$

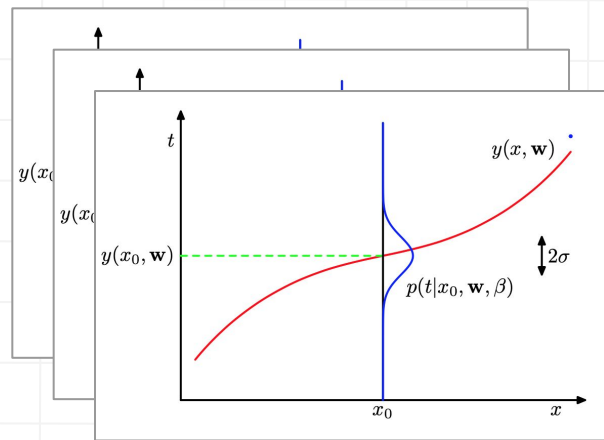
$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N) \quad (3.49)$$

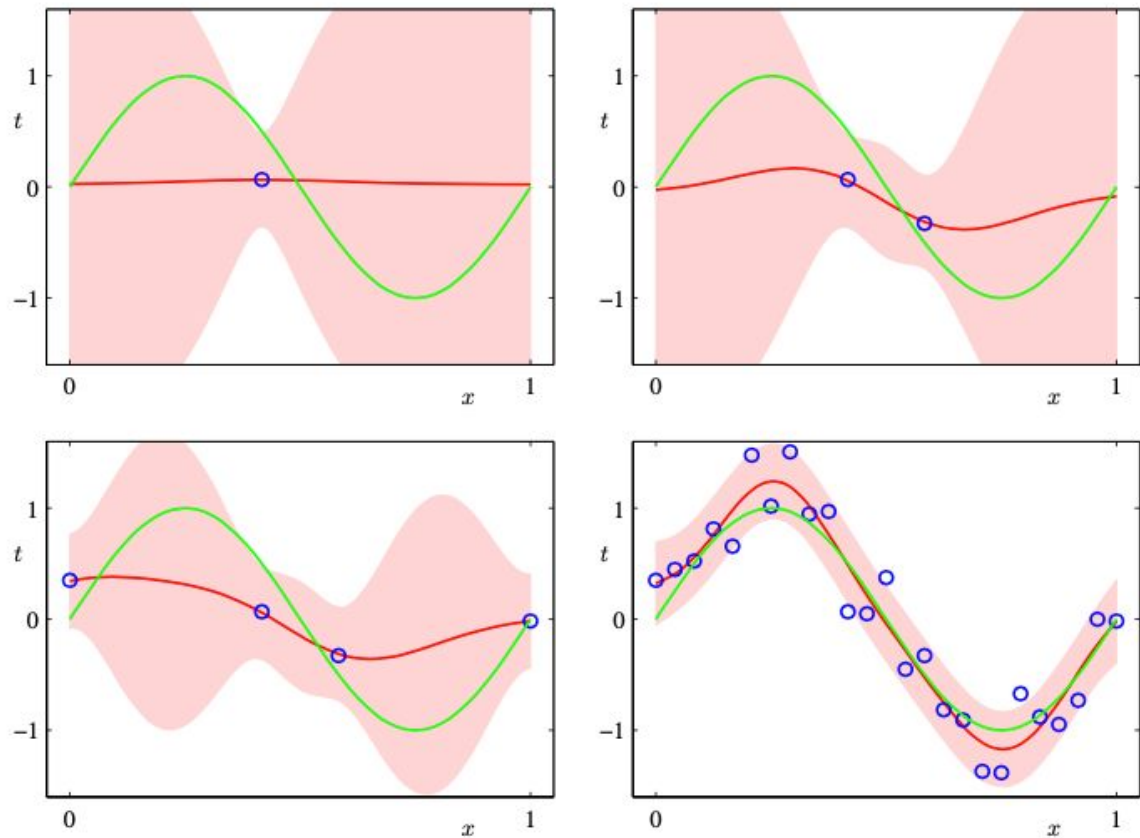
$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{L}^{-1} + \mathbf{A}\boldsymbol{\Lambda}^{-1}\mathbf{A}^T) \quad (2.115)$$

$$p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \quad (3.52)$$

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\phi}(\mathbf{x}), \sigma_N^2(\mathbf{x})) \quad (3.58)$$

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \boldsymbol{\phi}(\mathbf{x})^T \mathbf{S}_N \boldsymbol{\phi}(\mathbf{x}). \quad (3.59)$$

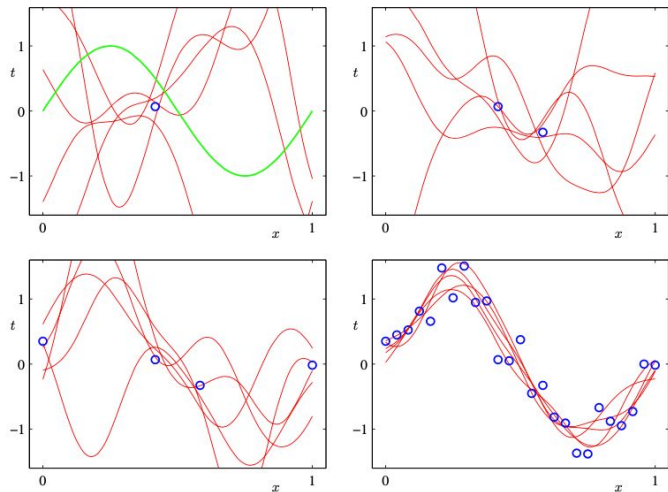




**Figure 3.8** Examples of the predictive distribution (3.58) for a model consisting of 9 Gaussian basis functions of the form (3.4) using the synthetic sinusoidal data set of Section 1.1. See the text for a detailed discussion.

$$p(t|\mathbf{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \phi(\mathbf{x}), \sigma_N^2(\mathbf{x}))$$

$$\sigma_N^2(\mathbf{x}) = \frac{1}{\beta} + \phi(\mathbf{x})^T \mathbf{S}_N \phi(\mathbf{x}).$$



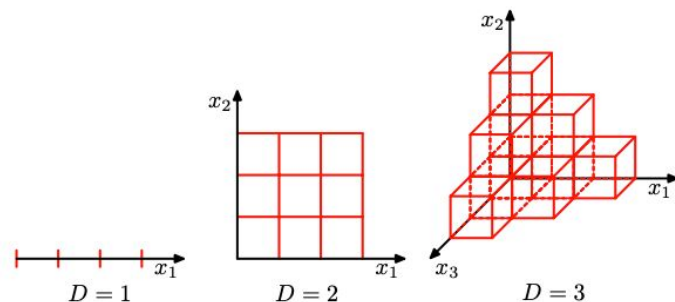
**Figure 3.9** Plots of the function  $y(x, \mathbf{w})$  using samples from the posterior distributions over  $\mathbf{w}$  corresponding to the plots in Figure 3.8.

$$p(\mathbf{w}|\mathbf{t}) = \mathcal{N}(\mathbf{w}|\mathbf{m}_N, \mathbf{S}_N)$$



# Curse of dimensionality

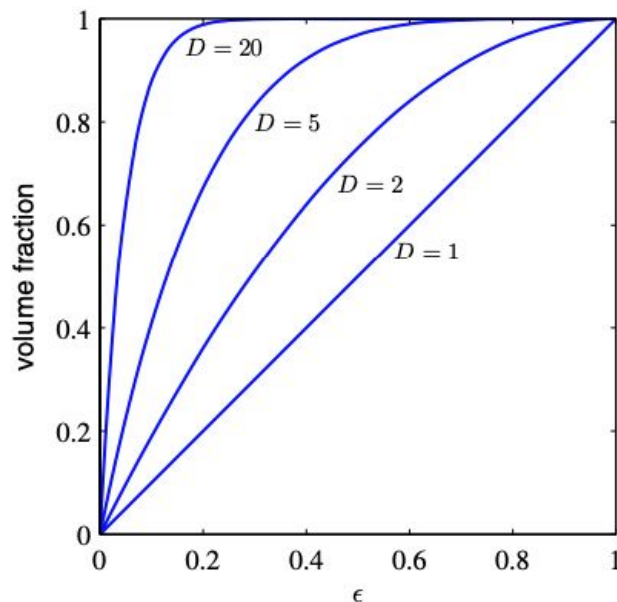
**Figure 1.21** Illustration of the curse of dimensionality, showing how the number of regions of a regular grid grows exponentially with the dimensionality  $D$  of the space. For clarity, only a subset of the cubical regions are shown for  $D = 3$ .



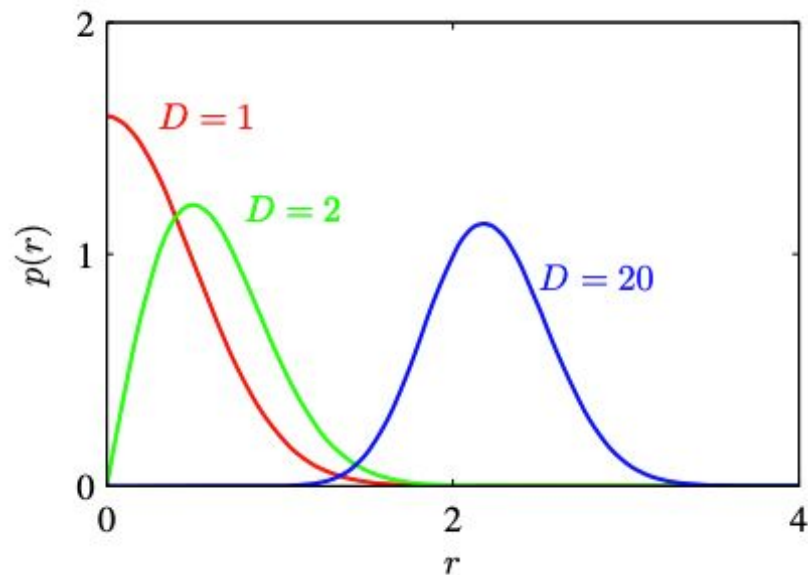
- Linear Algebra allows us to operate in  $n$ -dimensional vector spaces using the intuition from our 3-dimensional world as a vector space. No surprises as long as  $n$  is finite.
- If we add more structure to a vector space (e.g. inner product, metric), our intuition gained from the 3-dimensional world around us may be wrong.
- Example: Sphere of radius  $r = 1$ . What is the fraction of the volume of the sphere in a  $D$ -dimensional space which lies between radius  $r = 1$  and  $r = 1 - \epsilon$ ?
- Volume scales like  $r^D$ , therefore the formula for the volume of a sphere is  $V_D(r) = K_D r^D$ .

$$\frac{V_D(1) - V_D(1 - \epsilon)}{V_D(1)} = 1 - (1 - \epsilon)^D$$

Fig 1.22  
Eq (1.76)



**Figure 1.23** Plot of the probability density with respect to radius  $r$  of a Gaussian distribution for various values of the dimensionality  $D$ . In a high-dimensional space, most of the probability mass of a Gaussian is located within a thin shell at a specific radius.



$$x_1 = r \cos(\phi) \quad x_2 = r \sin(\phi)$$

$$p(r, \phi | 0, I) = \frac{1}{2\pi} r \exp \left\{ -\frac{1}{2} r^2 \right\}$$

# Discussions on fixed basis functions

**Pros:** assumption of linearity in the parameters (1) led to a range of useful properties including closed-form solutions to the least-squares problem, as well as a tractable Bayesian treatment. (2) for a suitable choice of basis functions, we can model arbitrary nonlinearities in the mapping from input variables to targets.

**Cons:** fixed basis functions before training data set is observed, the number of basis functions needs to grow rapidly, often exponentially, with the dimensionality  $D$  of the input space. Fixes: (1) localized basis functions that scatter only in regions containing data --> RBF networks, NN: adaptive basis functions. (2) target variables may have significant dependence on only a small number of possible directions. NN: choose response directions.

# What we have covered

Prelude: Conjugate prior, conditional of Gaussian variables, Bayes Theorem of Gaussian distributions

Bayesian regression

Predictive distributions

Curse of dimensionality

Info theory 101

# Information theory 101

How much information is a random variable  $x$ ?

the amount of information : ‘degree of surprise’ on learning the value of  $x$ ., say, expressed in function  $h(x)$

Assumptions:

- $h(x)$  is monotonic in the probability  $p(x)$
- For unrelated  $x \sim p(x)$ , and  $y \sim p(y)$ .  $h(x, y) = h(x) + h(y)$

(Information) Entropy [Shannon, 1948]

$$H[x] = - \sum_x p(x) \log_2 p(x). \quad (1.93)$$

# What is the unit of information entropy? - the origin of bits

Consider random variable  $x$ , with 8 equally likely states.  $H[x] = -8 \times \frac{1}{8} \log_2 \frac{1}{8} = 3 \text{ bits.}$

How to encode and transmit the value of  $x$  in binary digits? {000, 001, 010, 011, 100, 101, 110, 111}

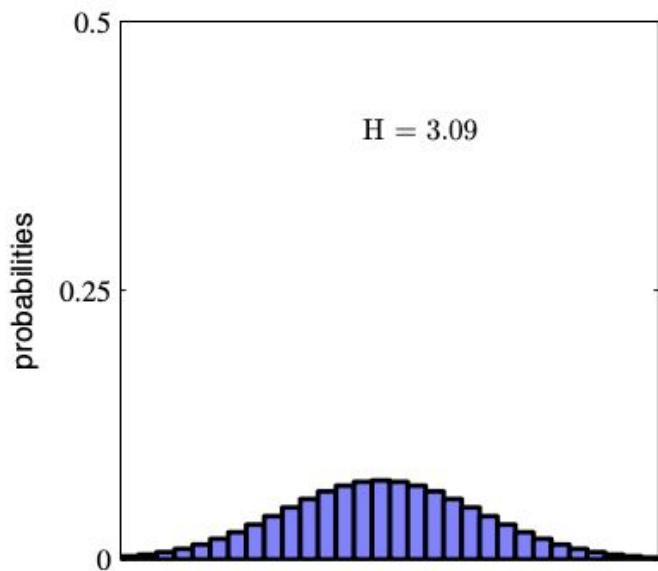
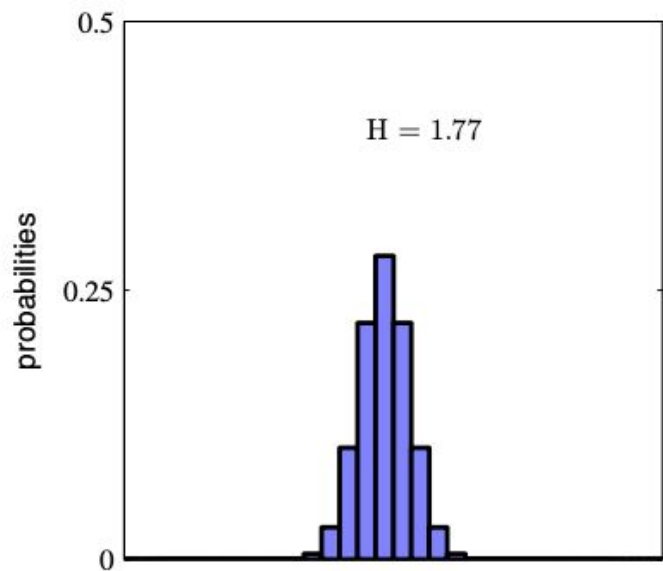
Now consider an example (Cover and Thomas, 1991) of a variable having 8 possible states  $\{a, b, c, d, e, f, g, h\}$  for which the respective probabilities are given by  $(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$ . The entropy in this case is given by

$$H[x] = -\frac{1}{2} \log_2 \frac{1}{2} - \frac{1}{4} \log_2 \frac{1}{4} - \frac{1}{8} \log_2 \frac{1}{8} - \frac{1}{16} \log_2 \frac{1}{16} - \frac{4}{64} \log_2 \frac{1}{64} = 2 \text{ bits.}$$

How to encode these? 0, 10, 110, 1110, 111100, 111101, 111110, 111111.

$$\text{average code length} = \frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{16} \times 4 + 4 \times \frac{1}{64} \times 6 = 2 \text{ bits}$$

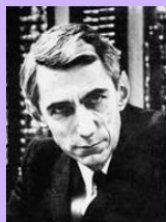
**Noiseless coding theorem** (Shannon 1948) the entropy is a lower bound on the number of bits needed to transmit the state of a random variable.



**Figure 1.30** Histograms of two probability distributions over 30 bins illustrating the higher value of the entropy  $H$  for the broader distribution. The largest entropy would arise from a uniform distribution that would give  $H = -\ln(1/30) = 3.40$ .

Reading: connection to entropy in physics

Recommended: “The Bit Player” (2018 documentary)



**Claude Shannon**  
1916–2001

After graduating from Michigan and MIT, Shannon joined the AT&T Bell Telephone laboratories in 1941. His paper ‘A Mathematical Theory of Communication’ published in the *Bell System Technical Journal* in 1948 laid the foundations for modern information the-

ory. This paper introduced the word ‘bit’, and his concept that information could be sent as a stream of 1s and 0s paved the way for the communications revolution. It is said that von Neumann recommended to Shannon that he use the term entropy, not only because of its similarity to the quantity used in physics, but also because “nobody knows what entropy really is, so in any discussion you will always have an advantage”.

# Entropy for continuous variables

$$H[\mathbf{x}] = - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x}. \quad (1.104)$$

What is a distribution  $p(x)$  that maximises  $H[x]$ , provided that  $p(x)$  is well defined, has given mean and variance? [Lagrange multipliers!](#)

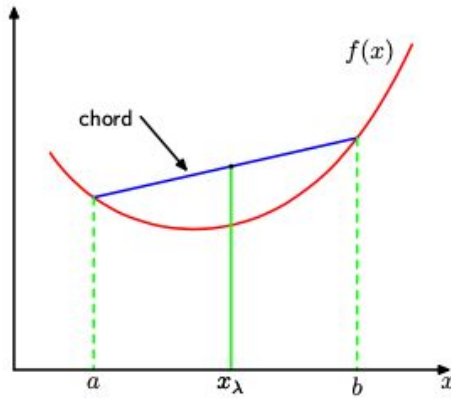
$$\begin{aligned} & - \int_{-\infty}^{\infty} p(x) \ln p(x) dx + \lambda_1 \left( \int_{-\infty}^{\infty} p(x) dx - 1 \right) \\ & + \lambda_2 \left( \int_{-\infty}^{\infty} xp(x) dx - \mu \right) + \lambda_3 \left( \int_{-\infty}^{\infty} (x - \mu)^2 p(x) dx - \sigma^2 \right) \end{aligned} \quad p(x) = \exp \left\{ -1 + \lambda_1 + \lambda_2 x + \lambda_3 (x - \mu)^2 \right\}. \quad (1.108)$$

$$p(x) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp \left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\} \quad (1.109)$$



# Convex functions

**Figure 1.31** A convex function  $f(x)$  is one for which every chord (shown in blue) lies on or above the function (shown in red).



$$f(\lambda a + (1 - \lambda)b) \leq \lambda f(a) + (1 - \lambda)f(b). \quad (1.114)$$

Jesen's inequality:  
 $f()$  - a convex function

$$f\left(\sum_{i=1}^M \lambda_i x_i\right) \leq \sum_{i=1}^M \lambda_i f(x_i)$$

where  $\lambda_i \geq 0$  and  $\sum_i \lambda_i = 1$ .

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

$$f\left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x}\right) \leq \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.$$

# KL divergence

$$\begin{aligned}\text{KL}(p\|q) &= - \int p(\mathbf{x}) \ln q(\mathbf{x}) d\mathbf{x} - \left( - \int p(\mathbf{x}) \ln p(\mathbf{x}) d\mathbf{x} \right) \\ &= - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x}.\end{aligned}\tag{1.113}$$

Apply Jensen's inequality,  $-\ln()$  is convex

$$f\left(\int \mathbf{x} p(\mathbf{x}) d\mathbf{x}\right) \leq \int f(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}.\tag{1.117}$$

$$\text{KL}(p\|q) = - \int p(\mathbf{x}) \ln \left\{ \frac{q(\mathbf{x})}{p(\mathbf{x})} \right\} d\mathbf{x} \geq - \ln \int q(\mathbf{x}) d\mathbf{x} = 0\tag{1.118}$$

Equality will only hold iff.  $p(x) = q(x)$  for all  $x$

If we have “incorrectly” represented  $p(x)$  with  $q(x)$ , how much more *information* do we need to recover  $p(x)$ ?

Not symmetric, not a distance measure!

But, has interesting algebraic and geometric properties, see Assignment 1

# “The Venn diagram of entropy and information”

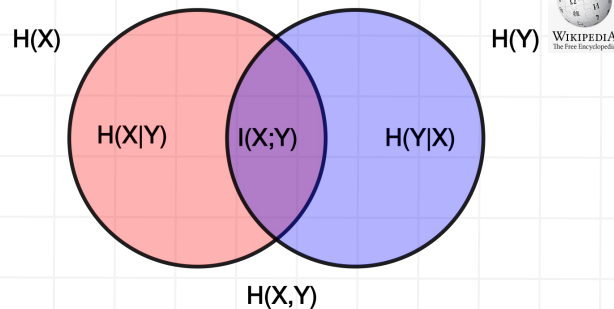
How much *information* is carried in  $x$  about  $y$ ? Two different metrics: conditional entropy and mutual information.

$$H[\mathbf{y}|\mathbf{x}] = - \iint p(\mathbf{y}, \mathbf{x}) \ln p(\mathbf{y}|\mathbf{x}) \, d\mathbf{y} \, d\mathbf{x} \quad (1.111)$$

$$H[\mathbf{x}, \mathbf{y}] = H[\mathbf{y}|\mathbf{x}] + H[\mathbf{x}] \quad (1.112)$$

$$\begin{aligned} I[\mathbf{x}, \mathbf{y}] &\equiv \text{KL}(p(\mathbf{x}, \mathbf{y}) \| p(\mathbf{x})p(\mathbf{y})) \\ &= - \iint p(\mathbf{x}, \mathbf{y}) \ln \left( \frac{p(\mathbf{x})p(\mathbf{y})}{p(\mathbf{x}, \mathbf{y})} \right) \, d\mathbf{x} \, d\mathbf{y} \end{aligned} \quad (1.120)$$

$$I[\mathbf{x}, \mathbf{y}] = H[\mathbf{x}] - H[\mathbf{x}|\mathbf{y}] = H[\mathbf{y}] - H[\mathbf{y}|\mathbf{x}]. \quad (1.121)$$



# Recap: Linear models for regression

- Basis functions
- Maximum Likelihood with Gaussian Noise
- Regularisation
- Bias variance decomposition

Prelude: Conjugate prior, conditional of Gaussian variables, Bayes Theorem of Gaussian distributions

Bayesian regression

Predictive distributions

Curse of dimensionality + discussions

Info theory 101