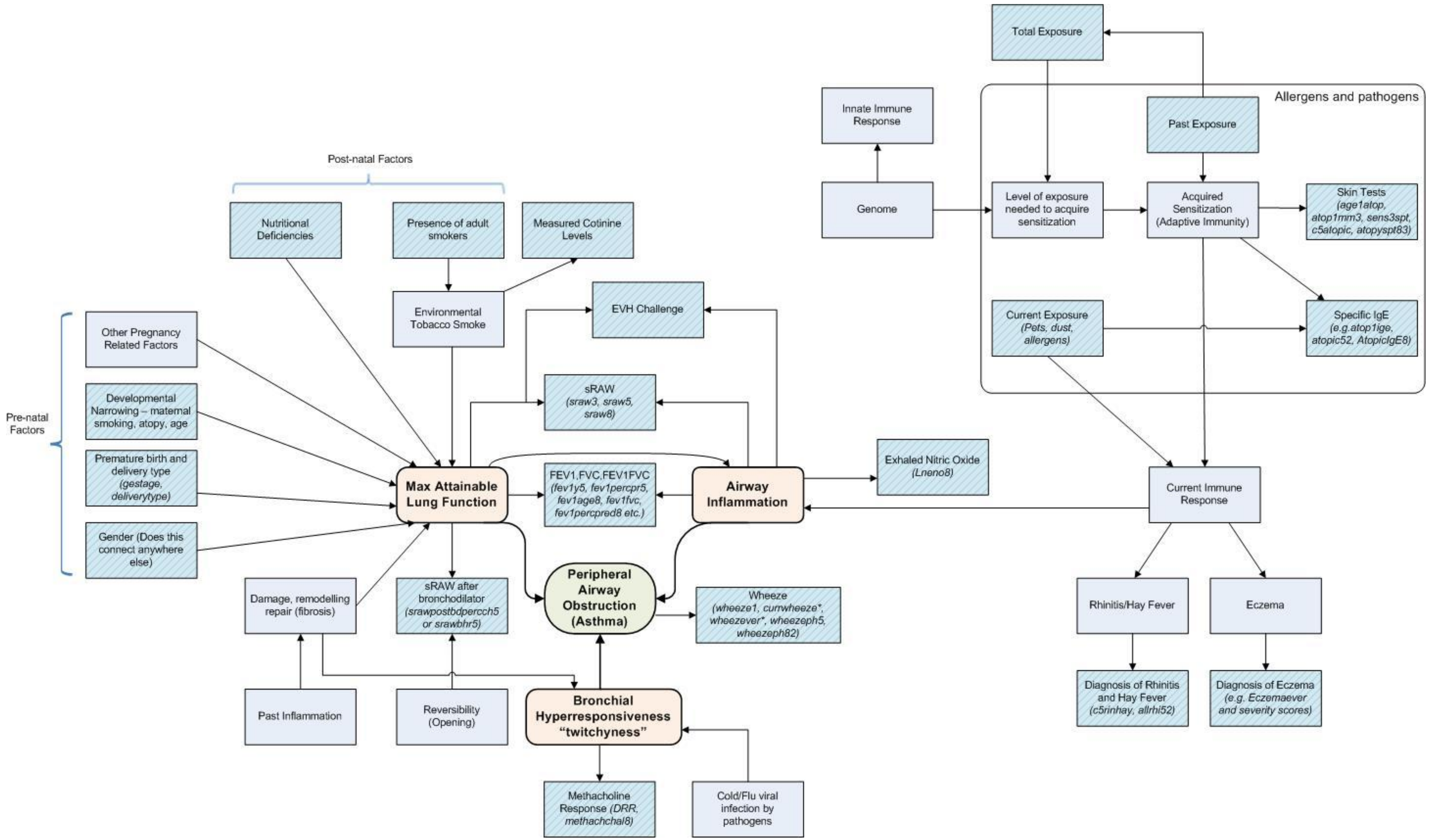
INTRODUCTION TO BAYESIAN INFERENCE – PART 2

CHRIS BISHOP



Post-natal Factors
Nutritional Deficiencies
Presence of adult smokers
Measured Cotinine Levels
Environmental Tobacco Smoke
EVH Challenge
Pre-natal Factors
Other Pregnancy Related Factors
Developmental Narrowing – maternal smoking, atopy, age
Premature birth and delivery type (gestage, deliverytype)
Gender (Does this connect anywhere else)
sRAW (sraw3, sraw5, sraw8)
Max Attainable Lung Function
FEV1,FVC,FEV1FVC (fev1y5, fev1percpr5, fev1age8, fev1fvc, fev1percpred8 etc.)
Airway Inflammation
Exhaled Nitric Oxide (Lneno8)
Peripheral Airway Obstruction (Asthma)
Wheeze (wheeze1, currwheeze*, wheezever*, wheezeph5, wheezeph82)
Damage, remodelling repair (fibrosis)
sRAW after bronchodilator (srawpostbdpercch5 or srawbhr5)
Past Inflammation
Reversibility (Opening)
Bronchial Hyperresponsiveness “twitchyness”
Methacholine Response (DRR, methachchal8)
Cold/Flu viral infection by pathogens
Total Exposure
Innate Immune Response
Allergens and pathogens
Past Exposure
Genome
Level of exposure needed to acquire sensitization
Acquired Sensitization (Adaptive Immunity)
Skin Tests (age1atop, atop1mm3, sens3spt, c5atopic, atopyspt83)
Current Exposure (Pets, dust, allergens)
Specific IgE (e.g.atop1ige, atopic52, AtopicIgE8)
Current Immune Response
Rhinitis/Hay Fever
Eczema
Diagnosis of Rhinitis and Hay Fever (c5rinhay, allrhi52)
Diagnosis of Eczema (e.g. Eczemaever and severity scores)

Personal Healthcare Revolution

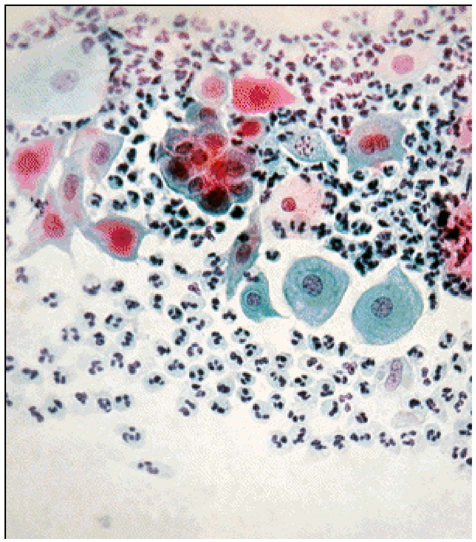
Electronic health records (CFH)

Personal genomics
  (DeCode, Navigenics, 23andMe)

X-prize: first $10k human genome technology
  NIH: $1k by 2014

Microsoft Research Cambridge:
  PhD Scholarships
  Internships: 3 months
  Postdoctoral Fellowships

Why Probabilities?



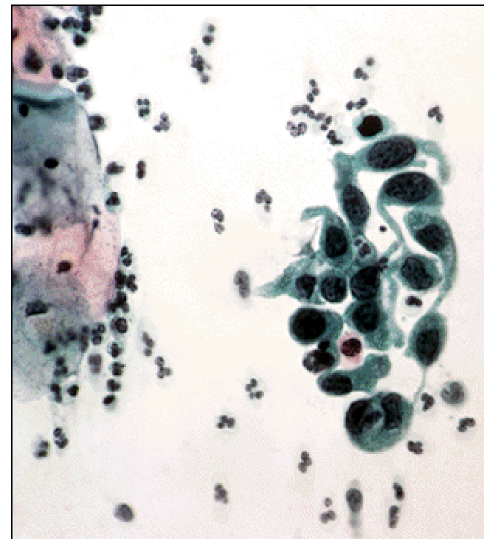


Image vector $\mathbf{x}$

Class $\mathcal{C}_k$ “cancer” or “normal”

Decisions

One-step solution

train a function to decide the class

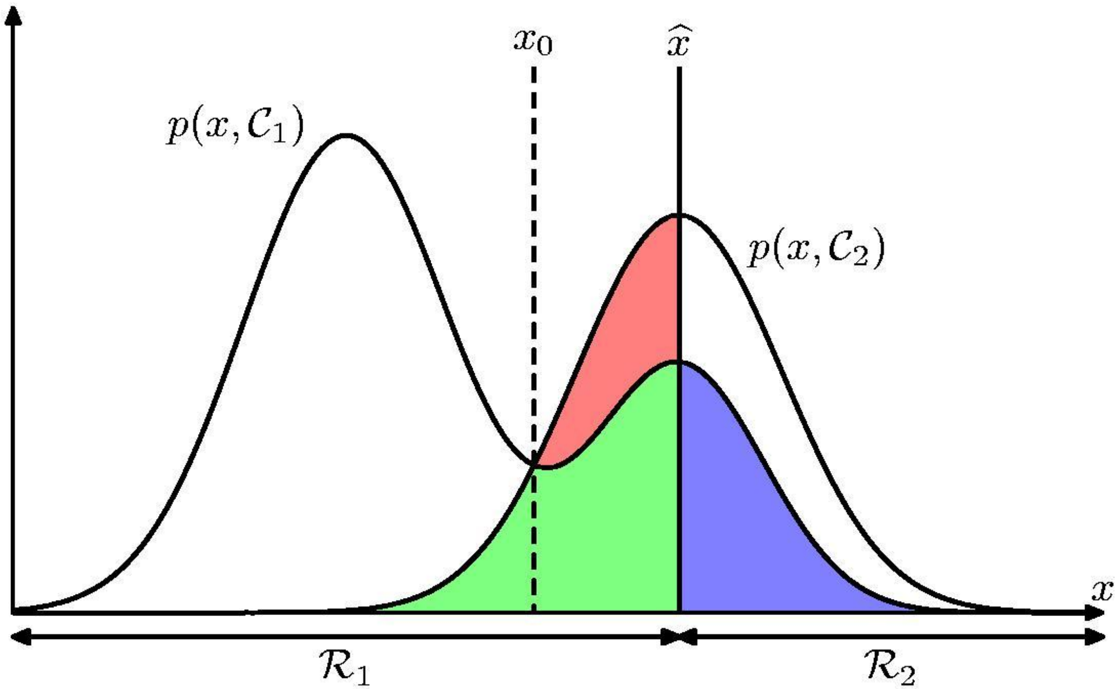
Two-step solution

inference : infer posterior probabilities

$$p(\mathcal{C}_k|\mathbf{x})$$

decision : use probabilities to decide the class

Minimum Misclassification Rate



$$p(\mathbf{x}, \mathcal{C}_1) > p(\mathbf{x}, \mathcal{C}_2)$$

$$p(\mathcal{C}_1|\mathbf{x}) > p(\mathcal{C}_2|\mathbf{x})$$

$$\begin{aligned} p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\ &= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\,\mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\,\mathrm{d}\mathbf{x}. \end{aligned}$$

Why Separate Inference and Decision?

- Minimizing risk (loss matrix may change over time)
- Reject option
- Unbalanced class priors
- Combining models

Loss Matrix

Decision

True class

$$\begin{array}{c} \\ \text{cancer} \\ \text{normal} \end{array} \begin{array}{c} \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \left(\begin{array}{cc} 0 & 1000 \\ 1 & 0 \end{array} \right) \end{array}$$

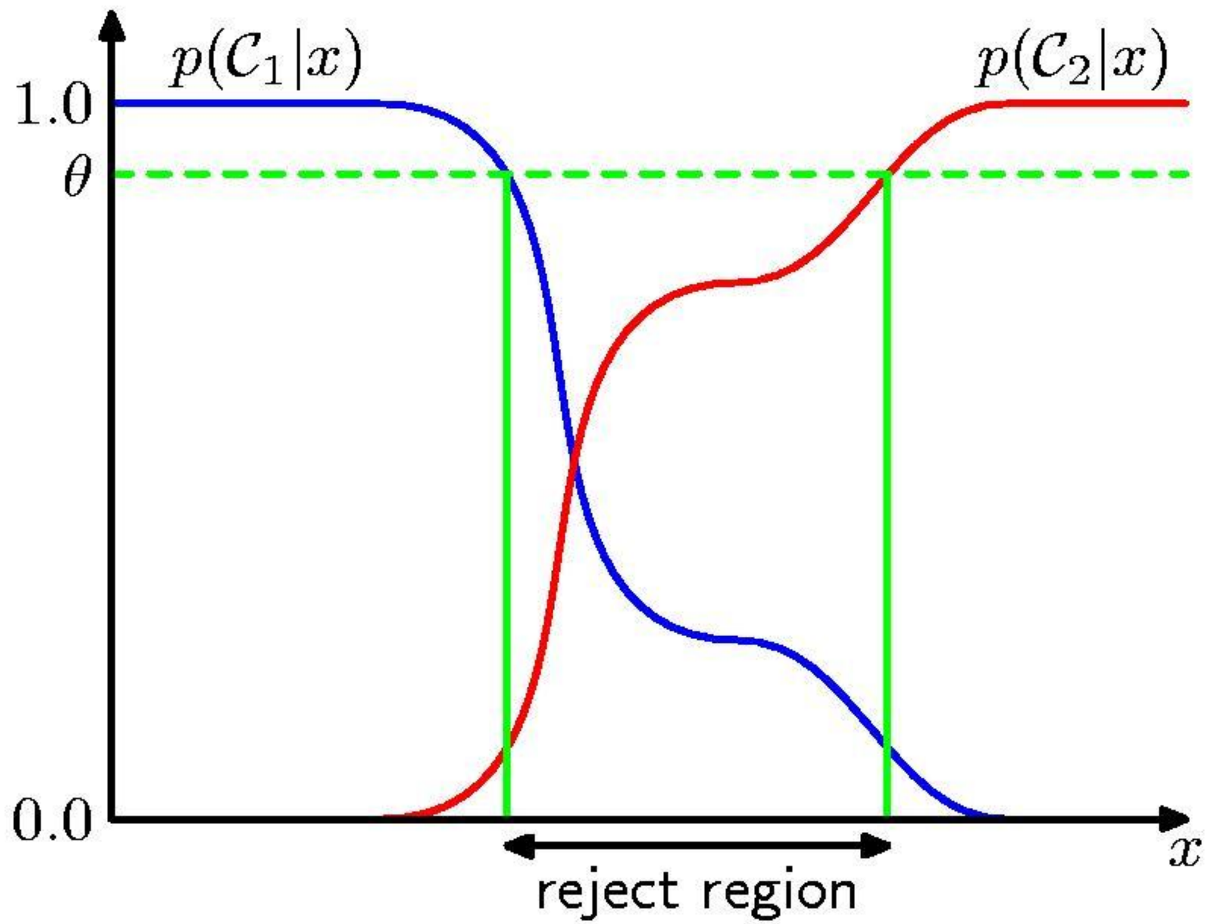
Minimum Expected Loss

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, \mathcal{C}_k) \, \mathrm{d}\mathbf{x}$$

Regions $\mathcal{R}_j$ are chosen, at each $\mathbf{x}$, to minimize

$$\sum_k L_{kj} p(\mathcal{C}_k | \mathbf{x})$$

Reject Option



Unbalanced class priors

In screening application, cancer is very rare

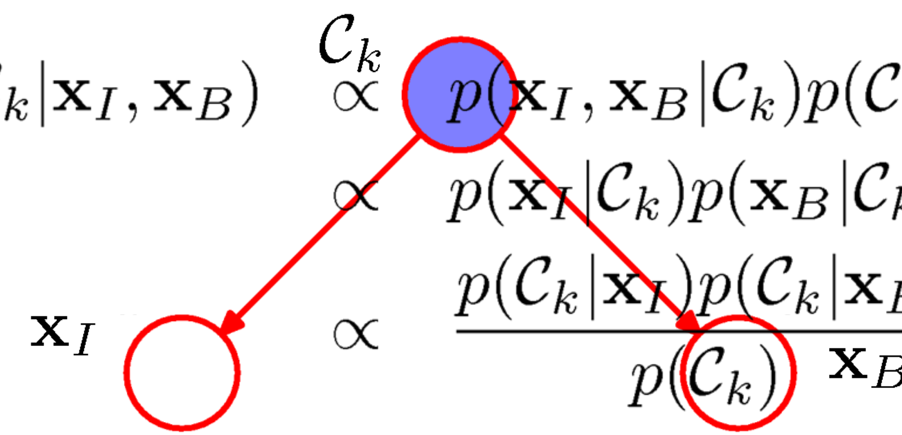
Use “balanced” data sets to train models, then use Bayes’ theorem to correct the posterior probabilities

Combining models

Image data and blood tests

Assume independent for each class:

$$p(\mathbf{x}_I, \mathbf{x}_B | \mathcal{C}_k) \propto p(\mathbf{x}_I | \mathcal{C}_k) p(\mathbf{x}_B | \mathcal{C}_k)$$



$$\begin{aligned} p(\mathcal{C}_k | \mathbf{x}_I, \mathbf{x}_B) &\propto p(\mathbf{x}_I, \mathbf{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) \\ &\propto p(\mathbf{x}_I | \mathcal{C}_k) p(\mathbf{x}_B | \mathcal{C}_k) p(\mathcal{C}_k) \\ &\propto \frac{p(\mathcal{C}_k | \mathbf{x}_I) p(\mathcal{C}_k | \mathbf{x}_B)}{p(\mathcal{C}_k)} \end{aligned}$$

Binary Variables (1)

Coin flipping: heads=1, tails=0

$$\begin{aligned} p(x=1|\mu) &= \mu \\ p(x=0|\mu) &= 1-\mu \end{aligned} \qquad \mu \in [0,1]$$

Bernoulli Distribution

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

Expectation and Variance

In general

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\, \mathrm{d}x$$

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

For Bernoulli

$$\begin{aligned} \operatorname{Bern}(x|\mu) &= \mu^x (1-\mu)^{1-x} \\ \mathbb{E}[x] &= \mu \\ \operatorname{var}[x] &= \mu(1-\mu) \end{aligned}$$

Likelihood function

Data set

$$\mathcal{D} = \{x_1, \ldots, x_N\}, \; m \text{ heads } (x=1), \; N-m \text{ tails } (x=0)$$

Likelihood function

$$\begin{aligned} p(\mathcal{D}|\mu) &= \prod_{n=1}^{N} p(x_n|\mu) \\ &= \prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n} \\ &= \mu^m (1-\mu)^{N-m} \end{aligned}$$

Prior Distribution

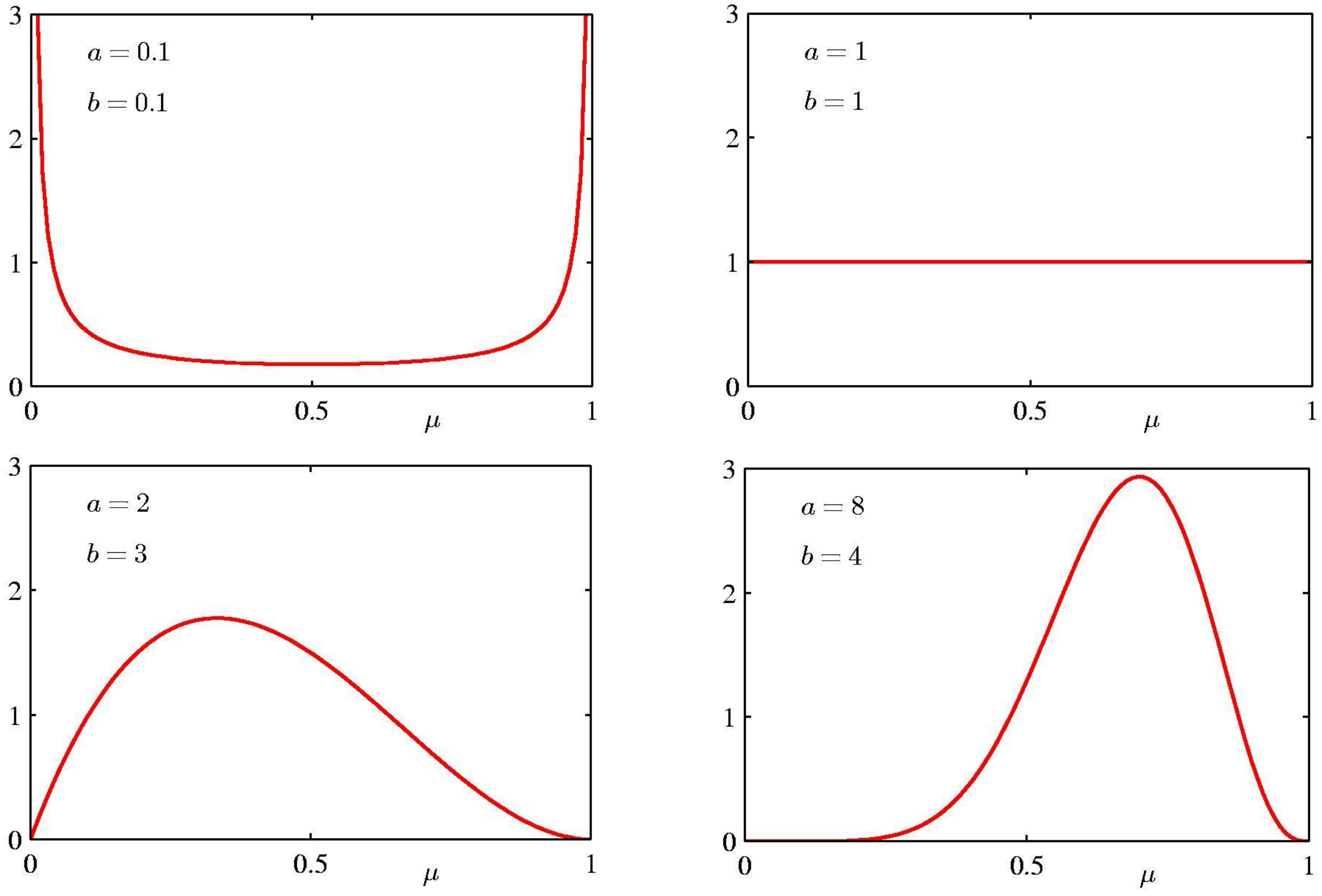
Simplification if prior has same functional form as likelihood function

$$p(\mu) \propto \mu^{a-1}(1-\mu)^{b-1}$$

Called *conjugate prior*

$$\begin{aligned}
\mathrm{Beta}(\mu|a,b) &= \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\mu^{a-1}(1-\mu)^{b-1} \\
\mathbb{E}[\mu] &= \frac{a}{a+b} \\
\mathrm{var}[\mu] &= \frac{ab}{(a+b)^2(a+b+1)}
\end{aligned}$$

Beta Distribution



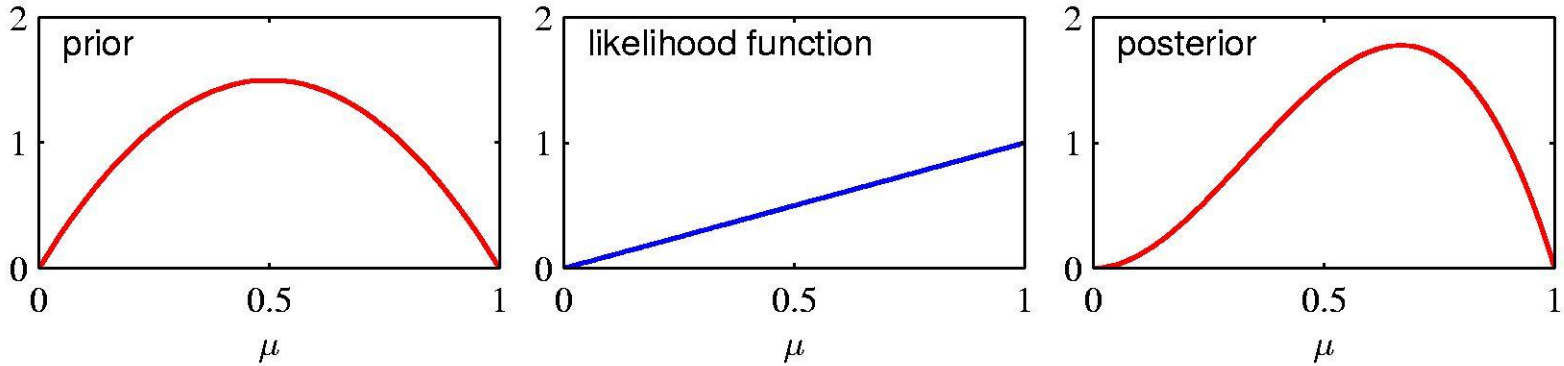
Posterior Distribution

$$\begin{aligned} p(\mu|a_0, b_0, \mathcal{D}) &\propto p(\mathcal{D}|\mu)p(\mu|a_0, b_0) \\ &\propto \left(\prod_{n=1}^{N} \mu^{x_n}(1-\mu)^{1-x_n}\right)\mathrm{Beta}(\mu|a_0, b_0) \\ &\propto \mu^{m+a_0-1}(1-\mu)^{(N-m)+b_0-1} \end{aligned}$$

$$p(\mu|a_0, b_0, \mathcal{D}) = \mathrm{Beta}(\mu|a_N, b_N)$$

$$a_N = a_0 + m \qquad b_N = b_0 + (N - m)$$

Posterior Distribution



Properties of the Posterior

As the size N of the data set increases

$$\begin{aligned}
a_N &\to m \\
b_N &\to N - m \\
\mathbb{E}[\mu] &= \frac{a_N}{a_N + b_N} \to \frac{m}{N} \\
\mathrm{var}[\mu] &= \frac{a_N b_N}{(a_N + b_N)^2 (a_N + b_N + 1)} \to 0
\end{aligned}$$

Predictive Distribution

What is the probability that the next coin flip will be heads?

$$\begin{aligned} p(x=1|a_0,b_0,\mathcal{D}) &= \int_0^1 p(x=1|\mu)p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\ &= \int_0^1 \mu p(\mu|a_0,b_0,\mathcal{D})\,\mathrm{d}\mu \\ &= \mathbb{E}[\mu|a_0,b_0,\mathcal{D}] \\ &= \frac{a_N}{a_N+b_N} \end{aligned}$$

The Exponential Family

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

where $\boldsymbol{\eta}$ is the *natural parameter*

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

We can interpret $g(\boldsymbol{\eta})$ as the normalization coefficient

Likelihood Function

Give a data set, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$p(\mathbf{X}|\boldsymbol{\eta}) = \left(\prod_{n=1}^{N} h(\mathbf{x}_n)\right) g(\boldsymbol{\eta})^N \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)\right\}$$

Depends on data through *sufficient statistics*

$$\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n)$$

Expected Sufficient Statistics

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

$$\nabla g(\boldsymbol{\eta}) \underbrace{\int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\}\, \mathrm{d}\mathbf{x}}_{1/g(\boldsymbol{\eta})} + \underbrace{g(\boldsymbol{\eta}) \int h(\mathbf{x}) \exp\left\{\boldsymbol{\eta}^{\mathrm{T}}\mathbf{u}(\mathbf{x})\right\} \mathbf{u}(\mathbf{x})\, \mathrm{d}\mathbf{x}}_{\mathbb{E}[\mathbf{u}(\mathbf{x})]} = 0$$

$$-\nabla \ln g(\boldsymbol{\eta}) = \mathbb{E}[\mathbf{u}(\mathbf{x})]$$

Conjugate priors

For the exponential family

$$p(\boldsymbol{\eta}|\boldsymbol{\chi}, \nu) = f(\boldsymbol{\chi}, \nu) g(\boldsymbol{\eta})^{\nu} \exp\left\{\nu \boldsymbol{\eta}^{\mathrm{T}} \boldsymbol{\chi}\right\}$$

Combining with the likelihood function, we get

$$p(\boldsymbol{\eta}|\mathbf{X}, \boldsymbol{\chi}, \nu) \propto g(\boldsymbol{\eta})^{\nu+N} \exp\left\{\boldsymbol{\eta}^{\mathrm{T}} \left(\sum_{n=1}^{N} \mathbf{u}(\mathbf{x}_n) + \nu \boldsymbol{\chi}\right)\right\}$$

Prior corresponds to ν pseudo-observations with statistic $\boldsymbol{\chi}$

Bernoulli revisited

The Bernoulli distribution

$$
\begin{aligned}
p(x|\mu) &= \mathrm{Bern}(x|\mu) = \mu^x (1-\mu)^{1-x} \\
&= \exp\{x \ln \mu + (1-x)\ln(1-\mu)\} \\
&= (1-\mu)\exp\left\{\ln\left(\frac{\mu}{1-\mu}\right)x\right\}
\end{aligned}
$$

Comparing with the general form we see that

$\eta = \ln\left(\frac{\mu}{1-\mu}\right)$ and so $\mu = \underbrace{\sigma(\eta) = \frac{1}{1+\exp(-\eta)}}_{\text{Logistic sigmoid}}$

Bernoulli revisited

The Bernoulli distribution in canonical form

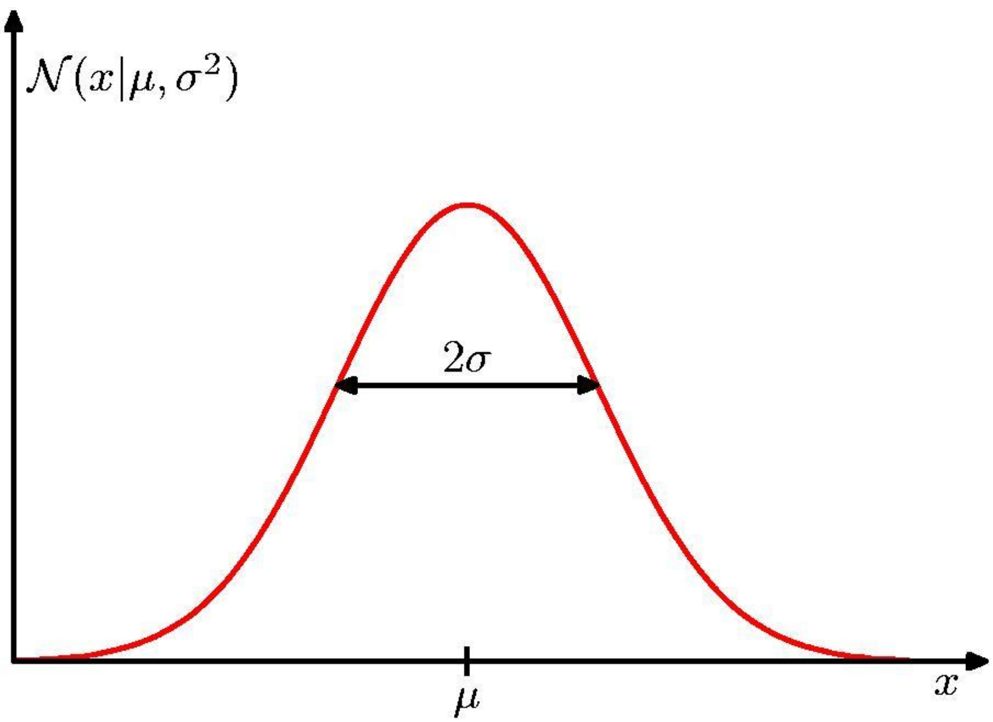
$$p(x|\eta) \quad = \quad h(x)g(\eta)\exp\left\{\eta^{\mathrm{T}}u(x)\right\}$$

where

$$\begin{aligned} u(x) &= x \\ h(x) &= 1 \\ g(\eta) &= 1 - \sigma(\eta) = \sigma(-\eta) \end{aligned}$$

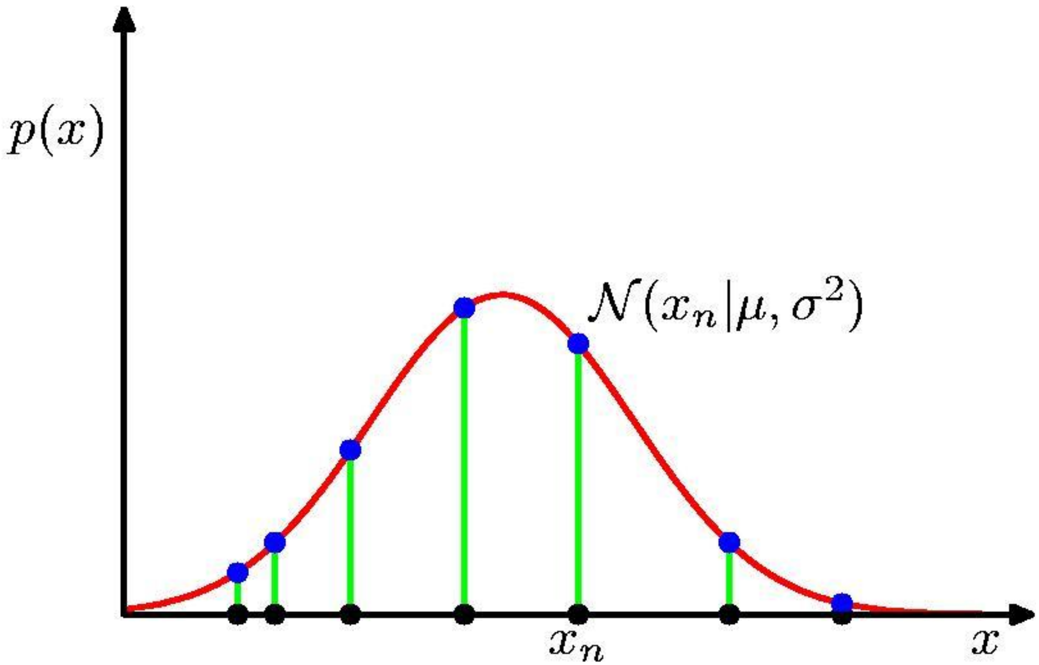
The Gaussian Distribution

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



Likelihood Function

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



Bayesian Inference – unknown mean

Assume σ^2 is known

Data set

$$\mathbf{x} = \{x_1, \ldots, x_N\}$$

Likelihood function for μ

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{ -\frac{1}{2\sigma^2} \sum_{n=1}^{N} (x_n - \mu)^2 \right\}.$$

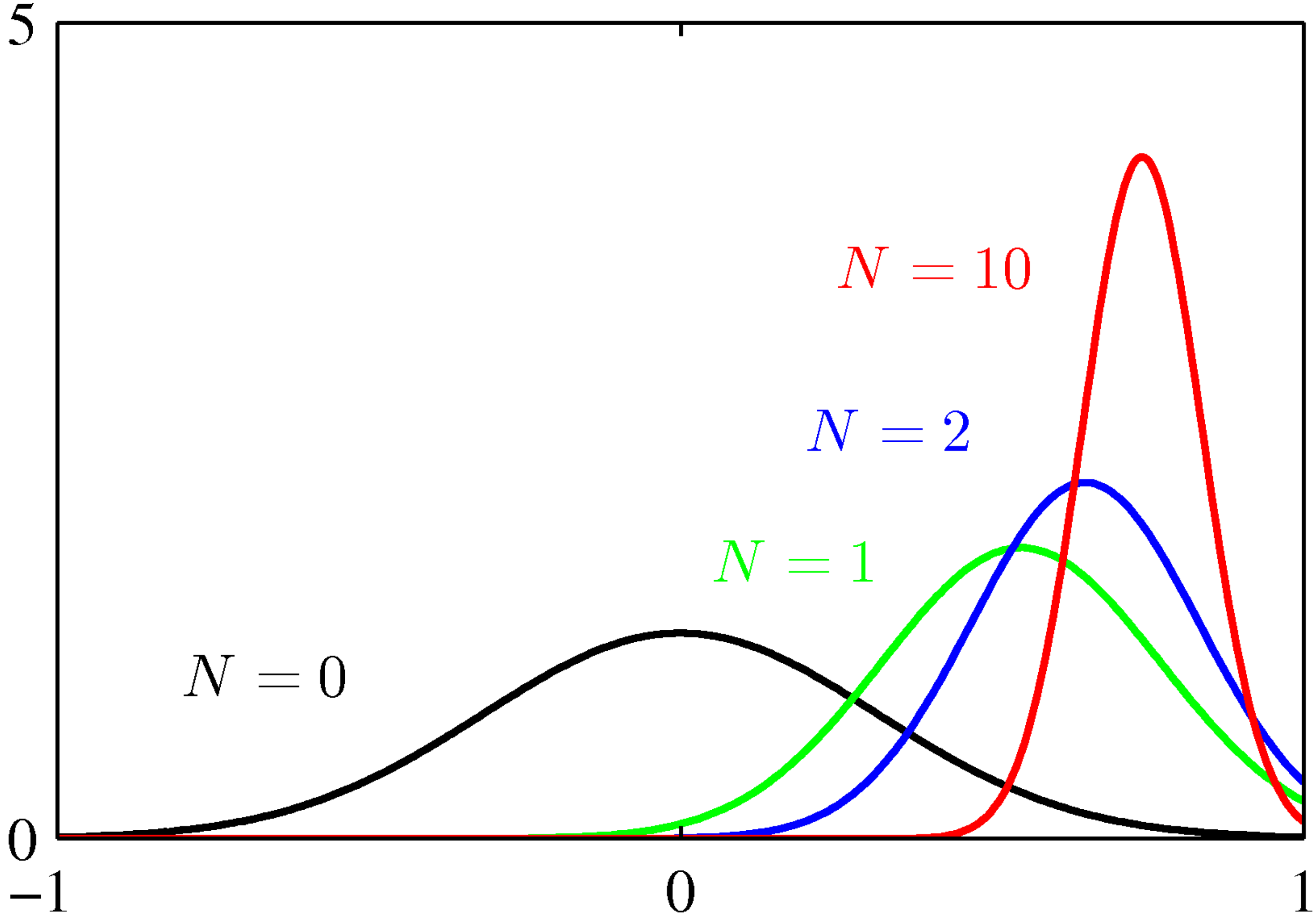
Bayesian Inference – unknown mean

Conjugate prior is a Gaussian

$$p(\mu) = \mathcal{N}\left(\mu | \mu_0, \sigma_0^2\right)$$

which gives a Gaussian posterior

$$p(\mu | \mathbf{x}) \propto p(\mathbf{x} | \mu) p(\mu)$$



5
0
−1
0
1
$N = 10$
$N = 2$
$N = 1$
$N = 0$

Bayesian Inference – unknown precision

Now assume μ is known

Likelihood function for precision $\lambda = 1/\sigma^2$

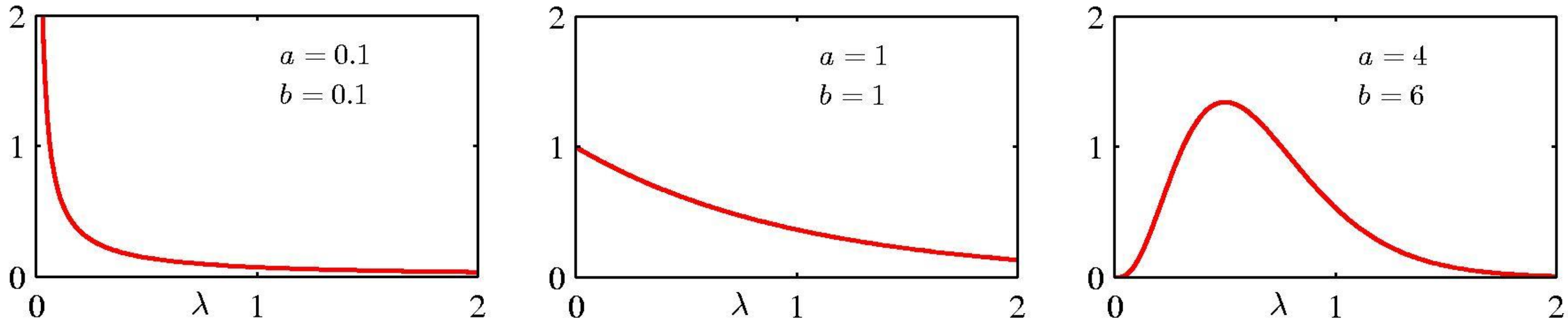
$$p(\mathbf{x}|\lambda) = \prod_{n=1}^{N} \mathcal{N}(x_n|\mu, \lambda^{-1}) \propto \lambda^{N/2} \exp\left\{-\frac{\lambda}{2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}$$

Conjugate prior

Gamma distribution

$$\mathrm{Gam}(\lambda|a,b) = \frac{1}{\Gamma(a)} b^a \lambda^{a-1} \exp(-b\lambda)$$

$$\mathbb{E}[\lambda] = \frac{a}{b} \qquad \mathrm{var}[\lambda] = \frac{a}{b^2}$$



Unknown Mean and Precision

Likelihood function

$$p(\mathbf{x}|\mu,\lambda) = \prod_{n=1}^{N}\left(\frac{\lambda}{2\pi}\right)^{1/2}\exp\left\{-\frac{\lambda}{2}(x_n-\mu)^2\right\}$$

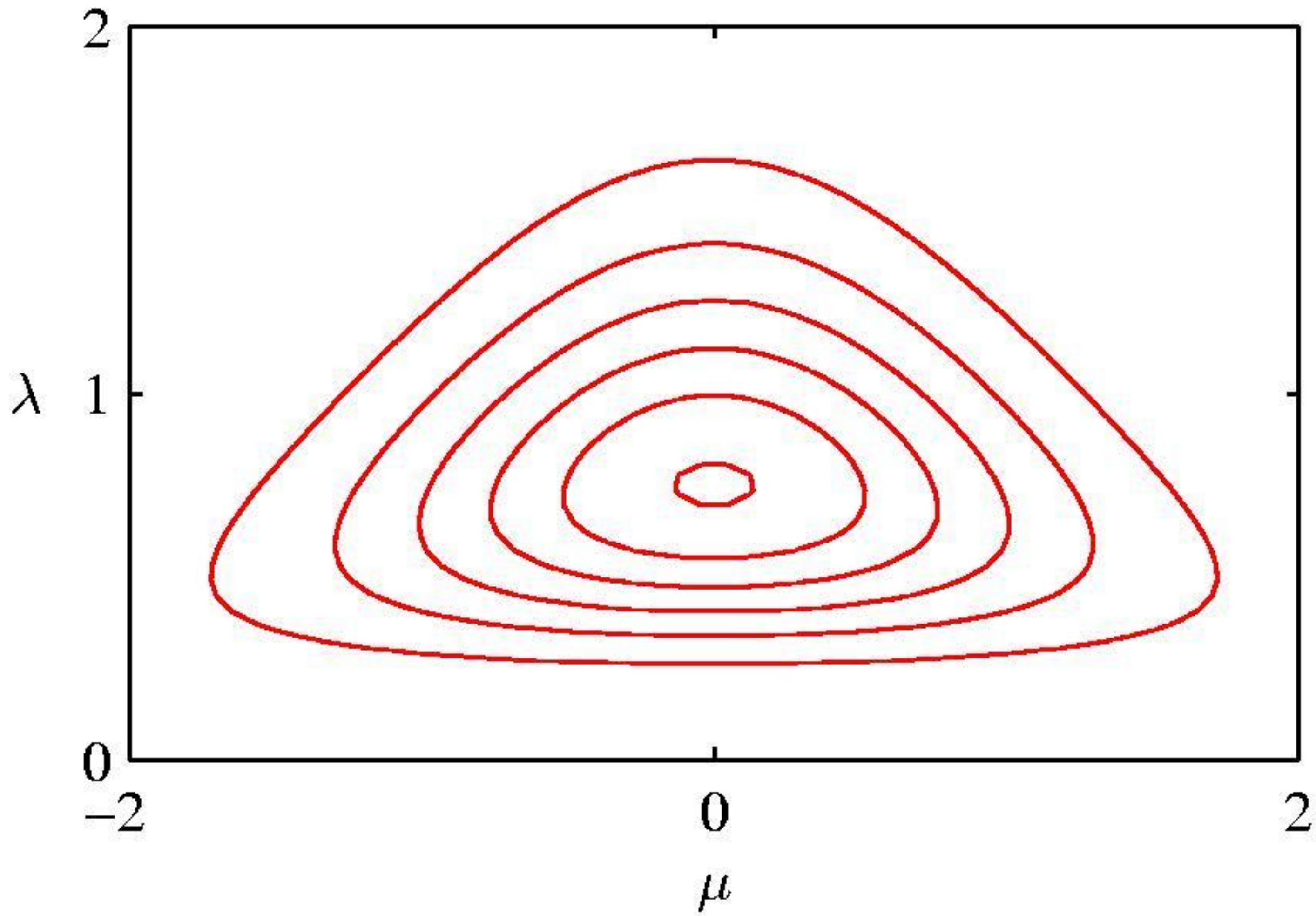
$$\propto \quad \left[\lambda^{1/2}\exp\left(-\frac{\lambda\mu^2}{2}\right)\right]^N \exp\left\{\lambda\mu\sum_{n=1}^{N}x_n - \frac{\lambda}{2}\sum_{n=1}^{N}x_n^2\right\}.$$

Gaussian-gamma distribution

$$p(\mu,\lambda) = p(\mu|\lambda)p(\lambda) = \mathcal{N}(\mu|\mu_0,(\beta\lambda)^{-1})\mathrm{Gam}(\lambda|a,b)$$

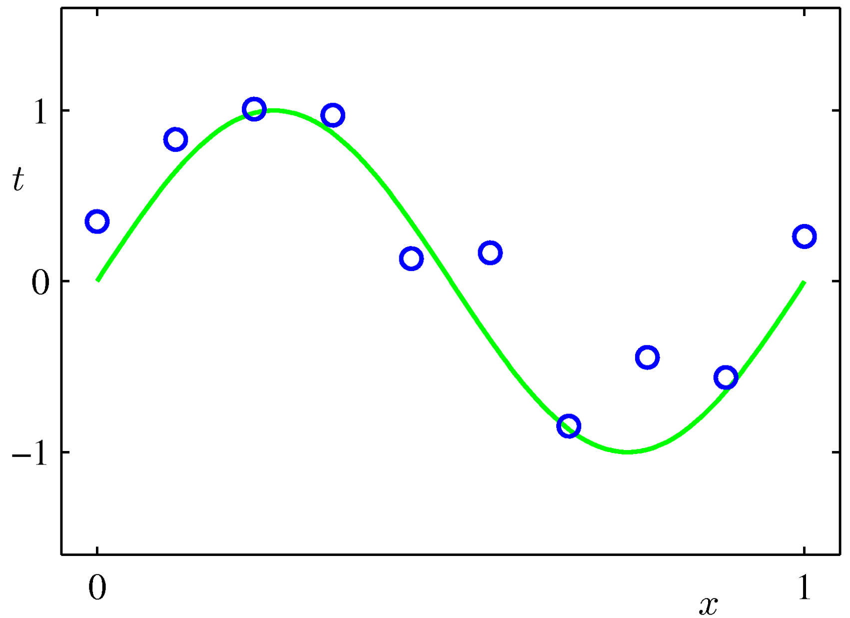
$$\propto \quad \exp\left\{-\frac{\beta\lambda}{2}(\mu-\mu_0)^2\right\}\lambda^{a-1}\exp\left\{-b\lambda\right\}$$

Gaussian-gamma Distribution



Linear Regression (1)

Noisy sinusoidal data



Linear Regression (2)

Linear combination of basis functions

$$y(\mathbf{x}, \mathbf{w}) = \sum_{j=0}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$$

Noise model

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

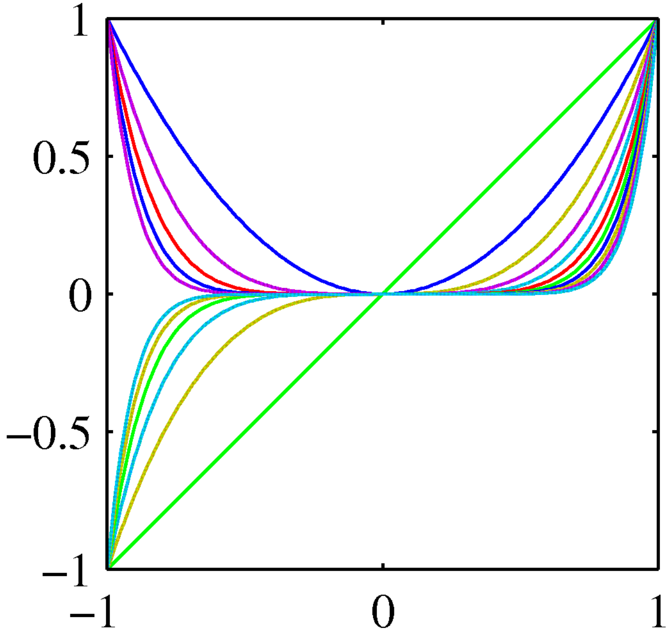
Likelihood function

$$p(\mathbf{t}|\mathbf{x}, \mathbf{w}, \beta) = \prod_{n=1}^{N} p(t_n|x_n, \mathbf{w}, \beta^{-1})$$

Linear Regression (3)

Polynomial basis functions

$$\phi_j(x) = x^j$$



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

Linear Regression (4)

Define a conjugate prior over $\mathbf{w}$

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

Combining with likelihood function gives the posterior

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

where

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

$$\Phi_{nj} = \phi_j(x_n)$$

Simple Example (1)

Data from straight line with Gaussian noise

$$\begin{aligned} t &= a + bx + \epsilon \\ \epsilon &\sim \mathcal{N}(\cdot|0,1) \end{aligned}$$

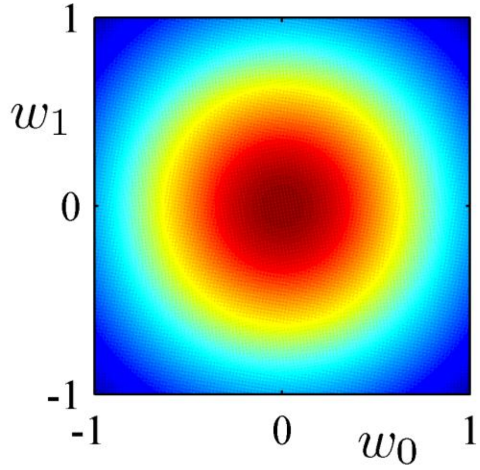
First order polynomial model

$$y(x, \mathbf{w}) = w_0 + w_1 x$$

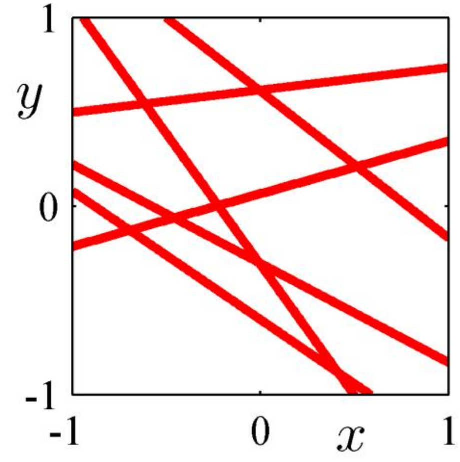
Simple Example (2)

0 data points observed

Prior

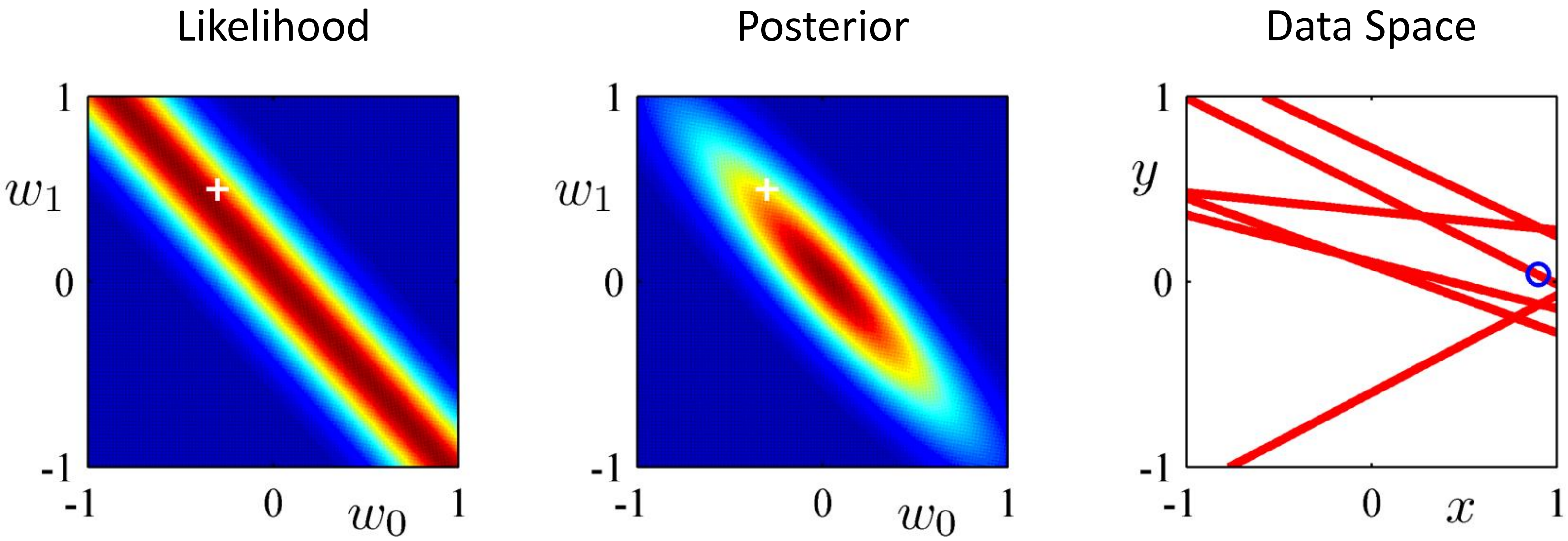


Data Space



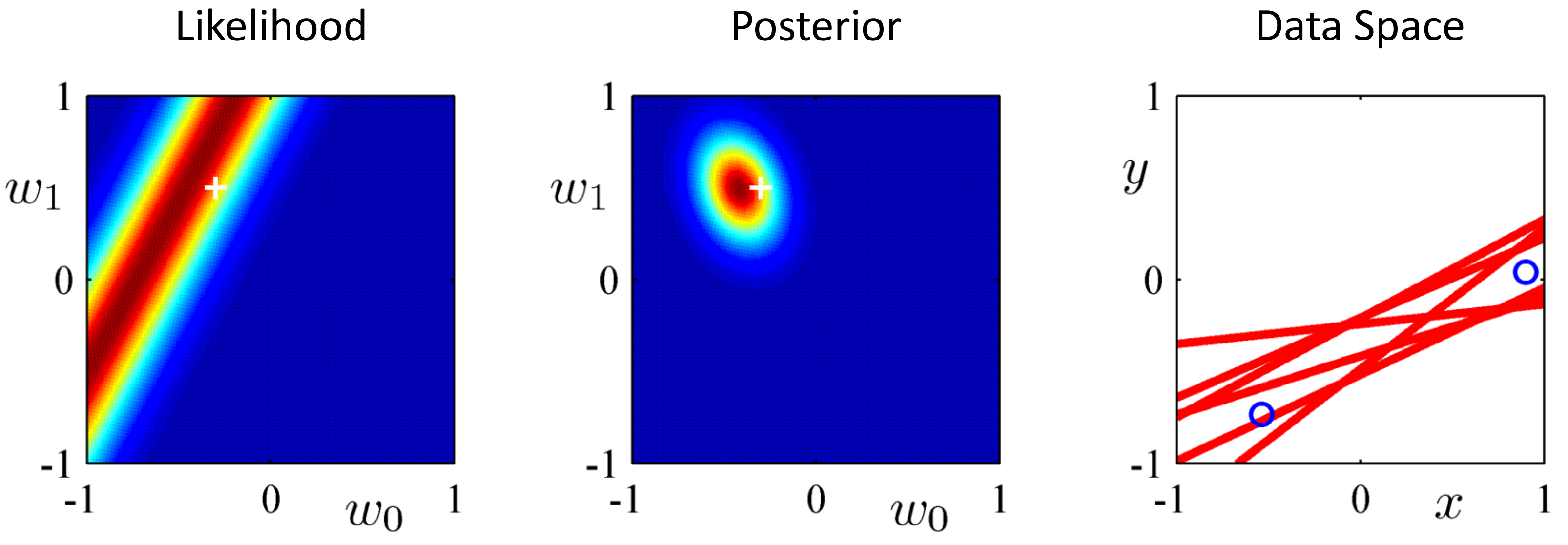
Simple Example (3)

1 data point observed



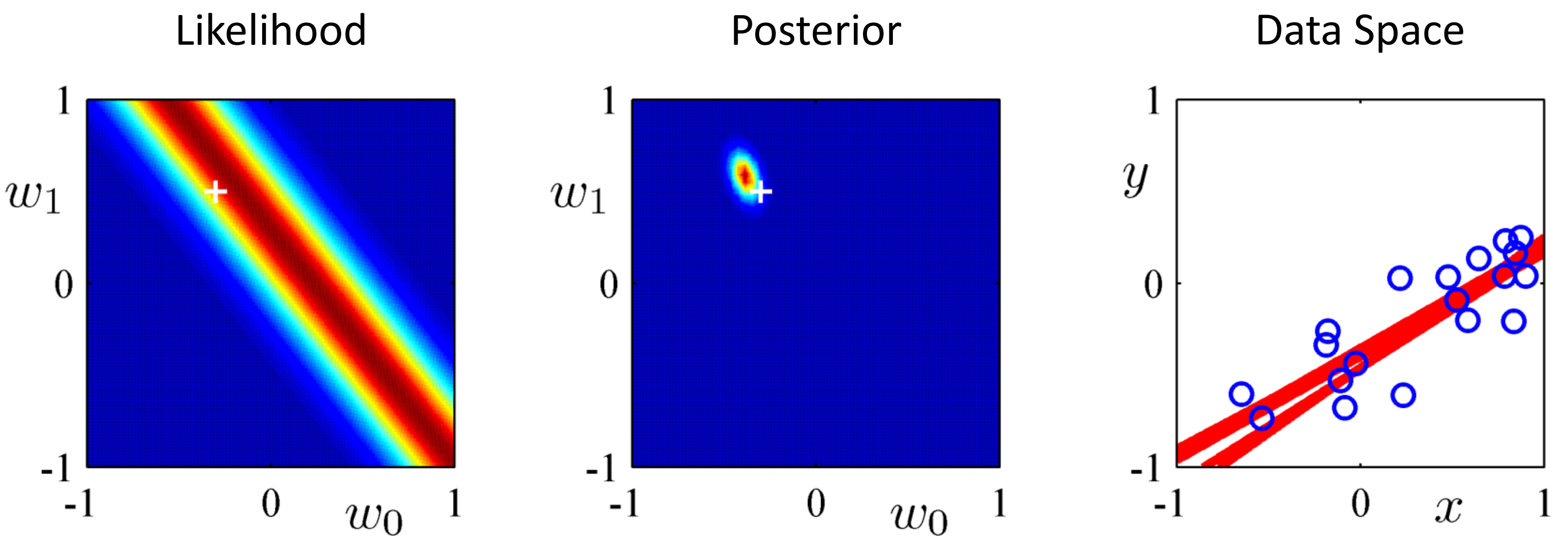
Simple Example (4)

2 data points observed



Simple Example (5)

20 data points observed



Predictive Distribution (1)

Predict t for new values of x by integrating over $\mathbf{w}$:

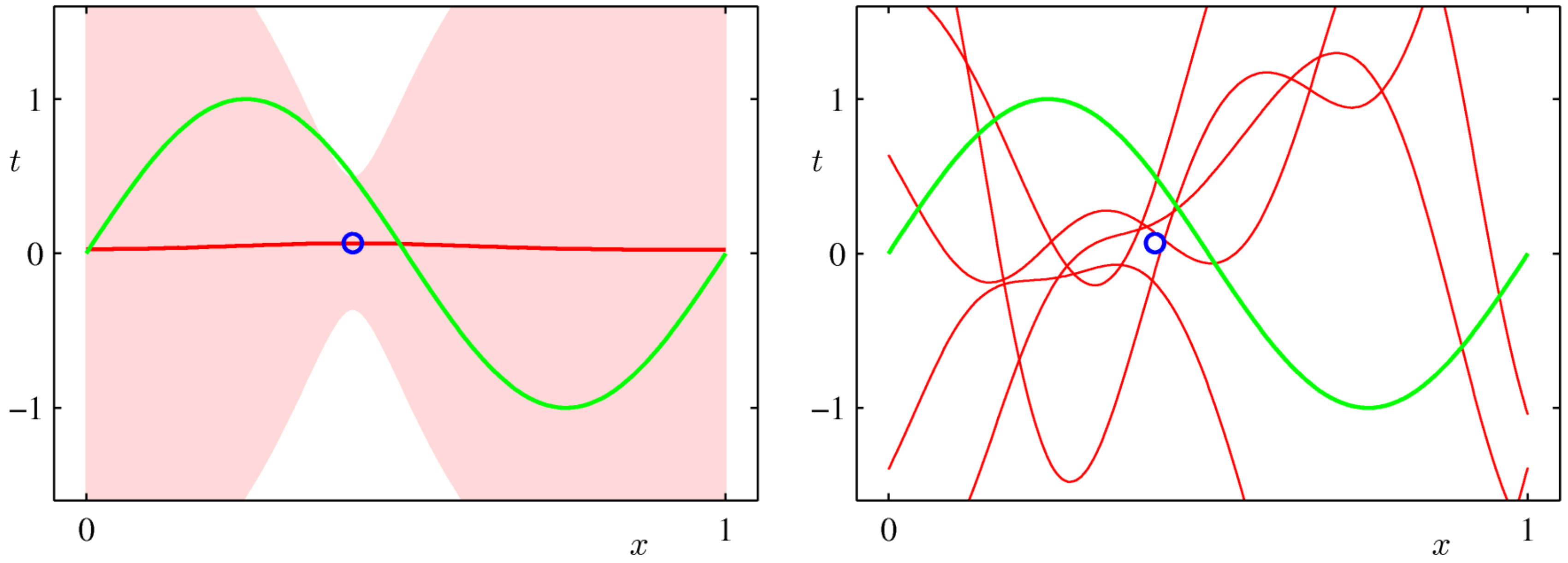
$$\begin{aligned} p(\widehat{t}|\mathbf{t}, \alpha, \beta, \widehat{x}) &= \int p(\widehat{t}|\mathbf{w}, \beta, \widehat{x}) p(\mathbf{w}|\mathbf{t}, \alpha, \beta) \, \mathrm{d}\mathbf{w} \\ &= \mathcal{N}\left(\widehat{t}|\mathbf{m}_N^{\mathrm{T}} \boldsymbol{\phi}(\widehat{x}), \sigma_N^2(\widehat{x})\right) \end{aligned}$$

where

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

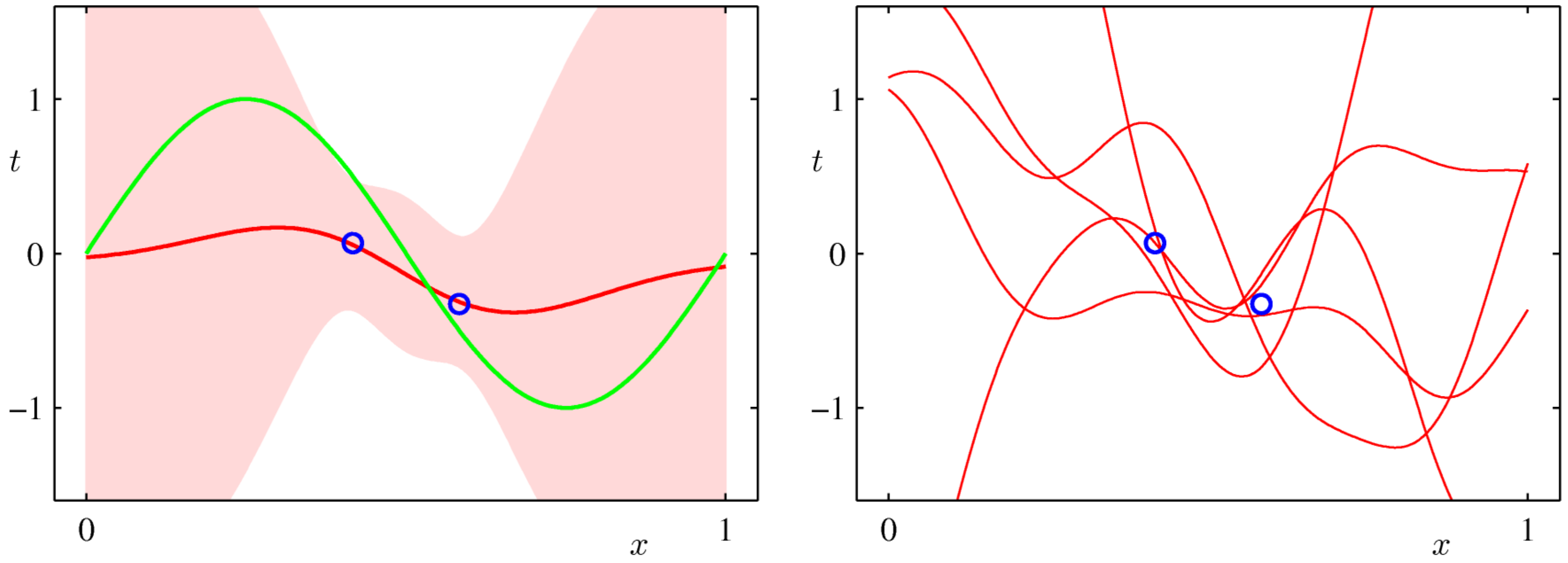
Predictive Distribution (3)

Example: Sinusoidal data, 9 Gaussian basis functions, 1 data point



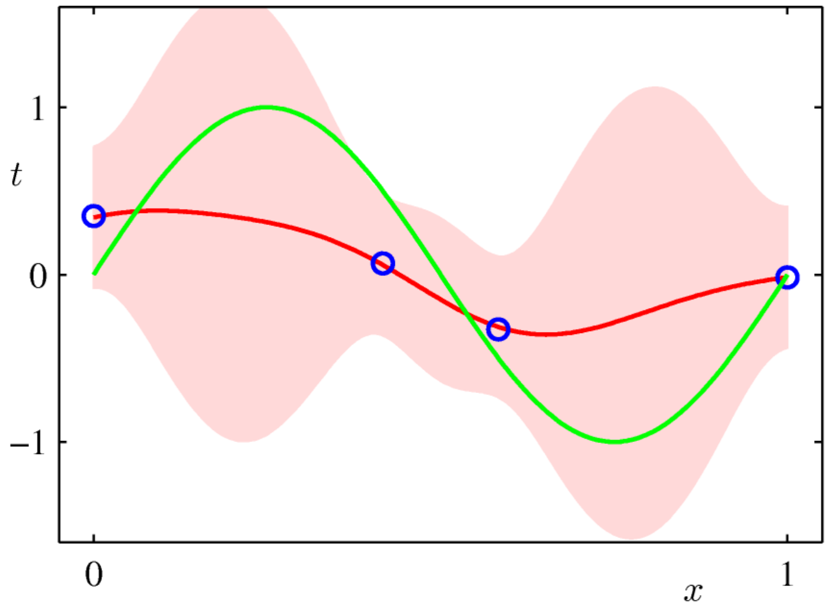
Predictive Distribution (4)

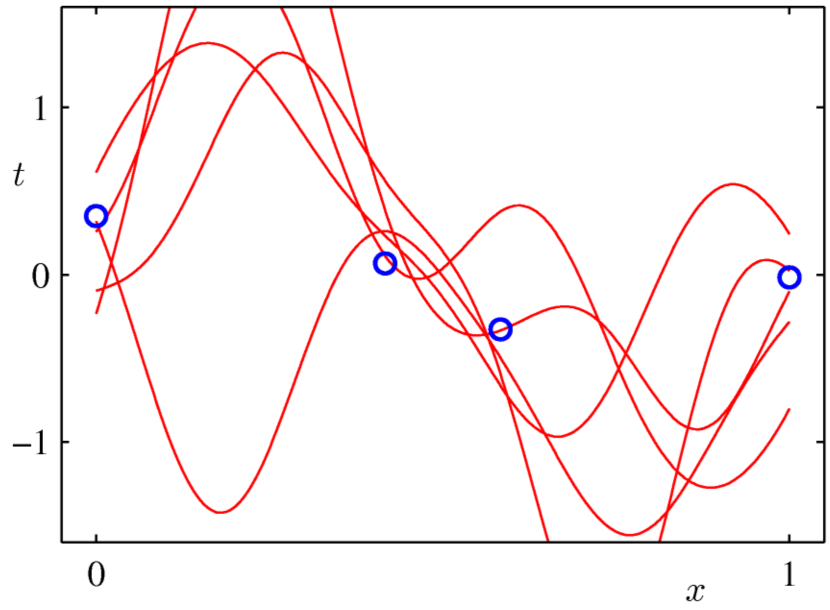
Example: Sinusoidal data, 9 Gaussian basis functions, 2 data points



Predictive Distribution (5)

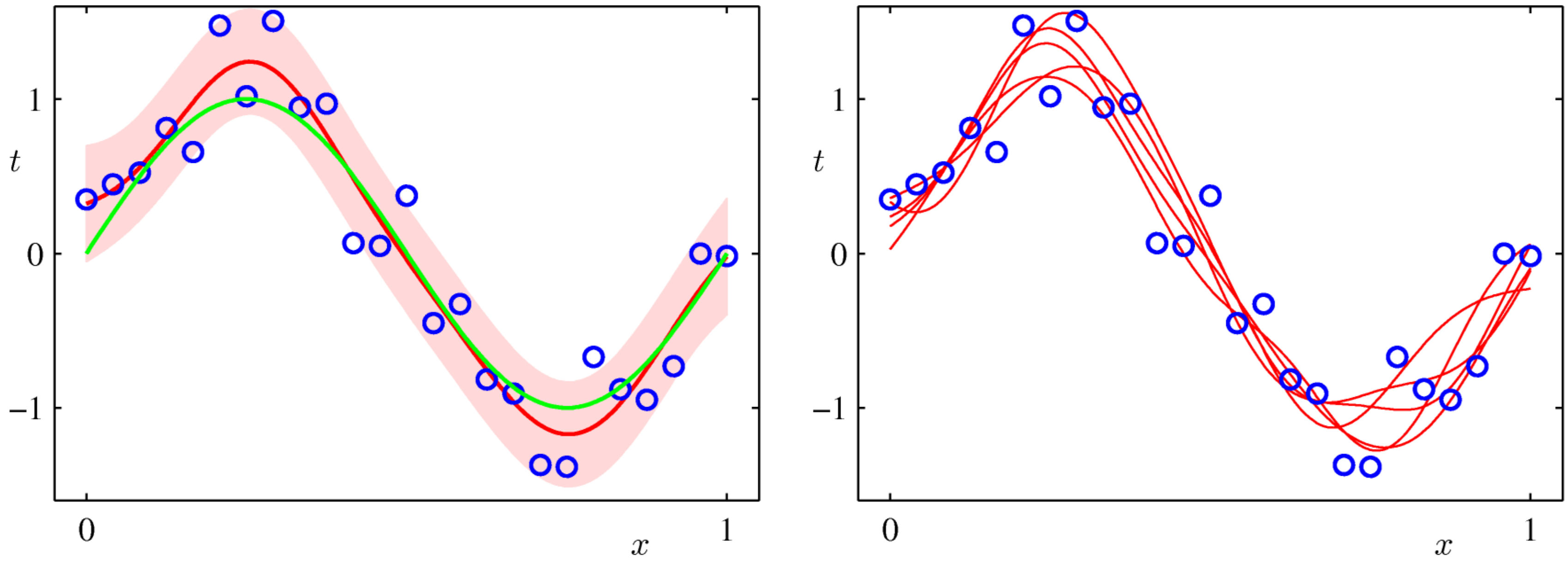
Example: Sinusoidal data, 9 Gaussian basis functions, 4 data points





Predictive Distribution (6)

Example: Sinusoidal data, 9 Gaussian basis functions, 25 data points



Bayesian Model Comparison (1)

Alternative models $\mathcal{M}_i,\ i{=}1,\ \ldots,L$

Predictive distribution is a mixture

$$p(t|\mathbf{x},\mathcal{D}) = \sum_{i=1}^{L} p(t|\mathbf{x},\mathcal{M}_i,\mathcal{D})p(\mathcal{M}_i|\mathcal{D})$$

Model selection: keep only most probable model

Bayesian Model Comparison (2)

From Bayes' theorem

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)$$

posterior

model evidence
(marginal likelihood)

prior

For equal priors, models ranked by marginal likelihood

Bayesian Model Comparison (4)

For a model with parameters $\mathbf{w}$

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i) \, \mathrm{d}\mathbf{w}$$

Note that

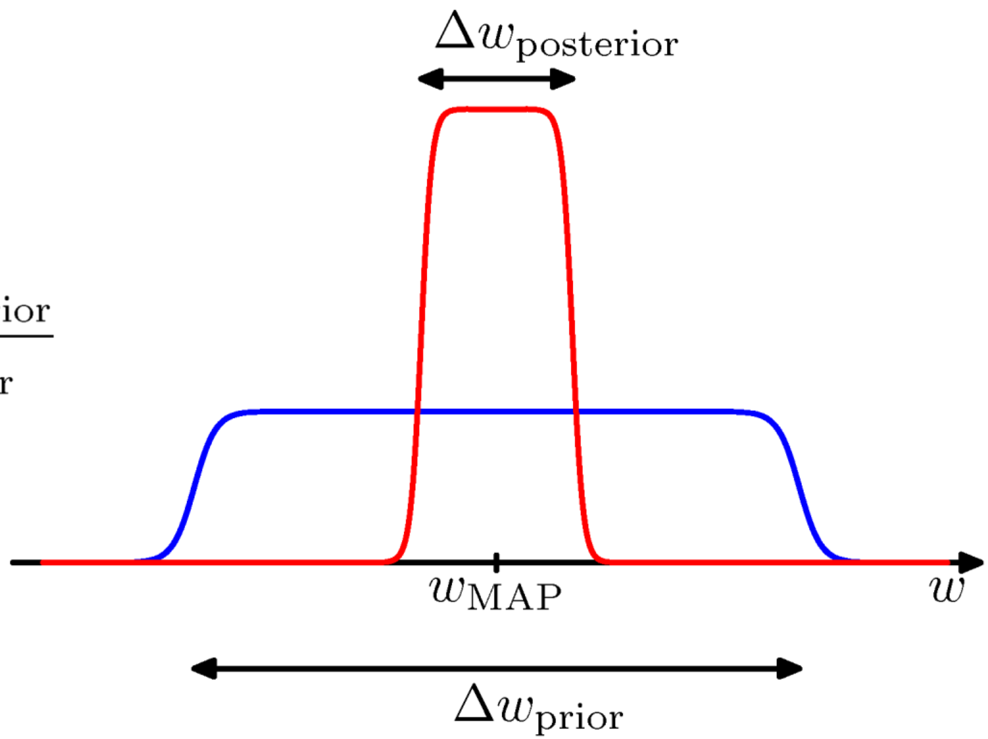
$$p(\mathbf{w}|\mathcal{D}, \mathcal{M}_i) = \frac{p(\mathcal{D}|\mathbf{w}, \mathcal{M}_i) p(\mathbf{w}|\mathcal{M}_i)}{p(\mathcal{D}|\mathcal{M}_i)}$$

Bayesian Model Comparison (5)

Consider model with a single parameter w

$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w)\,\mathrm{d}w$$

$$\simeq \quad p(\mathcal{D}|w_{\mathrm{MAP}})\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}$$



Bayesian Model Comparison (6)

Taking logarithms, we obtain

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{MAP}}) + \underbrace{\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)}_{\text{Negative}}$$

With M parameters, all assumed to have the same ratio $\Delta w_{\mathrm{posterior}}/\Delta w_{\mathrm{prior}}$, we get

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\mathrm{MAP}}) + M\ln\left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}}\right)$$

Linear Regression revisited

Marginal likelihood

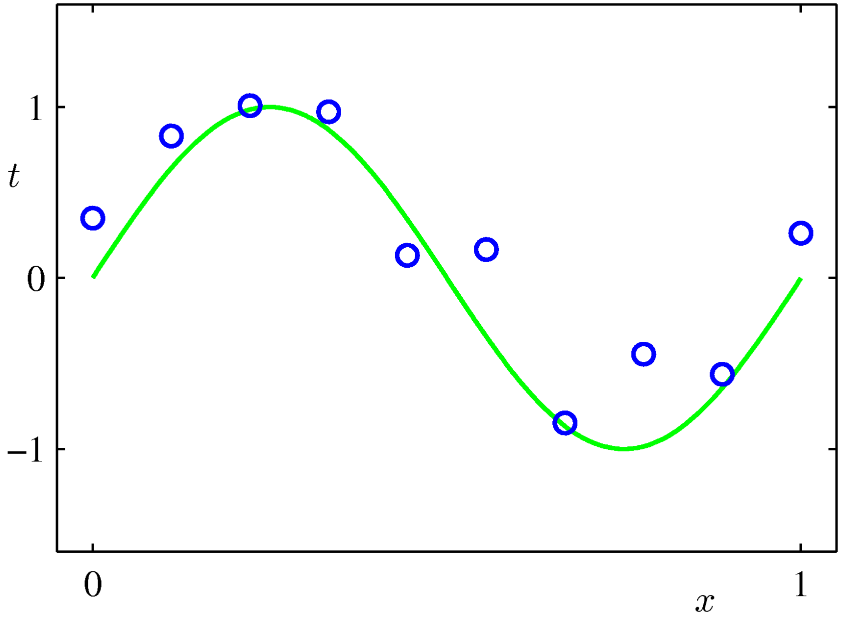
$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\mathbf{w},\beta)p(\mathbf{w}|\alpha)\,\mathrm{d}\mathbf{w}$$

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) + \frac{1}{2}\ln|\mathbf{S}_N| - \frac{N}{2}\ln(2\pi)$$

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

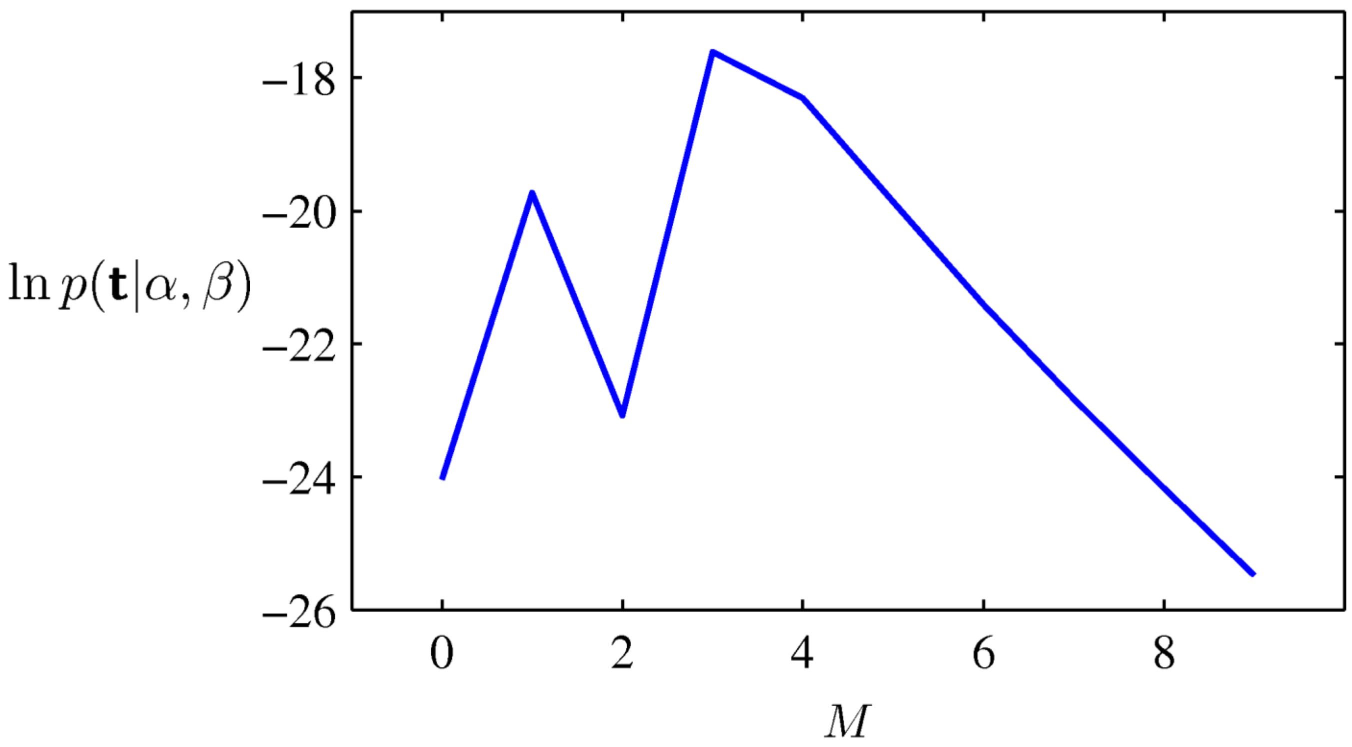
Linear Regression revisited

Noisy sinusoidal data



Linear Regression revisited

Polynomial of order M, $\alpha = 5 \times 10^{-3}$



Bayesian Model Comparison

Matching data and model complexity