

# **ECE521:**

## **Lectures 12 & 13**

27 February & 2 March 2017:

PCA continued,  
Bayesian methods

*With thanks to Russ Salakhutdinov and others*

# This week

- We will explore the Bayesian framework further, including Bayesian Linear Regression Models
- Examples of additional perspectives:
  - Bishop 2006: section 3.3
  - Murphy 2012: parts of chap. 5, & sec. 7.6
- On Thursday, midterms will be distributed as well

# Bayesian Approach

- We formulate our knowledge about the world probabilistically:
  - We **define the model** that expresses our knowledge qualitatively (e.g. independence assumptions, forms of distributions).
  - Our model will have some **unknown parameters**.
  - We capture our assumptions, or prior beliefs, about unknown parameters (e.g. range of plausible values) by **specifying the prior distribution** over those parameters before seeing the data.
- We **observe the data**.
- We compute the **posterior probability distribution** for the parameters, given observed data.
- We use this posterior distribution to:
  - **Make predictions** by averaging over the posterior distribution
  - **Examine/Account for uncertainty** in the parameter values.
  - **Make decisions** by minimizing expected posterior loss.

# Posterior Distribution

- The posterior distribution for the model parameters can be found by combining the prior with the likelihood for the parameters given the data.
- This is accomplished using **Bayes' Rule**:

$$P(\text{parameters} \mid \text{data}) = \frac{P(\text{data} \mid \text{parameters})P(\text{parameters})}{P(\text{data})}$$

Probability of  
observed data  
given  $\mathbf{w}$

Prior probability of  
weight vector  $\mathbf{w}$

$$p(\mathbf{w} \mid \mathcal{D}) = \frac{p(\mathcal{D} \mid \mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}$$

Posterior probability  
of weight vector  $\mathbf{W}$   
given training data  $\mathcal{D}$

Marginal likelihood  
(normalizing constant):

$$P(\mathcal{D}) = \int p(\mathcal{D} \mid \mathbf{w})P(\mathbf{w})d\mathbf{w}$$

This integral can be high-dimensional and is often difficult to compute.

# The Rules of Probability

Sum Rule:

$$p(X) = \sum_Y p(X, Y)$$

Product Rule:

$$p(X, Y) = p(Y|X)p(X)$$

# Predictive Distribution

- We can also state Bayes' rule in words:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

- We can make predictions for a new data point  $\mathbf{x}^*$ , given the training dataset by **integrating over the posterior distribution**:

$$p(\mathbf{x}^*|\mathcal{D}) = \int p(\mathbf{x}^*|\mathbf{w}, \mathcal{D})p(\mathbf{w}|\mathcal{D})d\mathbf{w} = \mathbb{E}_{P(\mathbf{w}|\mathcal{D})}[p(\mathbf{x}^*|\mathbf{w}, \mathcal{D})],$$

which is sometimes called the **predictive distribution**.

- Note that computing the predictive distribution requires knowledge of the posterior distribution:

$$p(\mathbf{w}|\mathcal{D}) = \frac{p(\mathcal{D}|\mathbf{w})P(\mathbf{w})}{P(\mathcal{D})}, \quad \text{where } P(\mathcal{D}) = \int p(\mathcal{D}|\mathbf{w})P(\mathbf{w})d\mathbf{w}$$

which is usually **intractable**.

# Modelling Challenges

- The first challenge is in specifying suitable model and suitable prior distributions. This can be challenging particularly when dealing with high-dimensional problems we see in machine learning.
  - A suitable model should admit all the possibilities that are thought to be at all likely.
  - A suitable prior should avoid giving zero or very small probabilities to possible events, but should also avoid spreading out the probability over all possibilities.
- We may need to properly model dependencies among parameters in order to avoid having a prior that is too spread out.
- One strategy is to introduce latent variables into the model and hyperparameters into the prior.
- Both of these represent the ways of modelling dependencies in a tractable way.

# Computational Challenges

The other big challenge is computing the posterior distribution. There are several main approaches:

- **Analytical integration**: If we use “conjugate” priors, the posterior distribution can be computed analytically. Chiefly employed for simple models


- 
- **Gaussian (Laplace) approximation**: Approximate the posterior distribution with a Gaussian. Works well when there is a lot of data compared to the model complexity (as posterior is close to Gaussian).

(We aren't covering the last three here)

- **Monte Carlo integration**: Once we have a sample from the posterior distribution, we can do many things. The dominant current approach is Markov Chain Monte Carlo (MCMC): simulate a Markov chain that converges to the posterior distribution. It can be applied to a wide variety of problems.
- **Variational approximation**: A cleverer way to approximate the posterior. It often works much faster compared to MCMC. But often not as general as MCMC.



# Our linear regression techniques

- LLS LR = MLE LR:  $\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} p(D|\mathbf{w})$
- MAP LR:  $\hat{\mathbf{w}} = \operatorname{argmax}_{\mathbf{w}} p(\mathbf{w}|D)$ 
  - $\ell_2$  regularization combats overfitting
  - $\ell_1$  regularization does so with sparser solutions 
- Bayesian LR:  $p(\mathbf{w}|D)$ 
  - Combats overfitting while allowing more data to be used for training
  - N.B.: something called “Empirical-Bayes LR” (not covered here) reduces the assumptions we make about the prior

# Bayesian Linear Regression

- Given observed inputs  $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N\}$ , and corresponding target values  $\mathbf{t} = [t_1, t_2, \dots, t_N]^T$ , we can write down the likelihood function:

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

where  $\boldsymbol{\phi}(\mathbf{x}) = (\phi_0(\mathbf{x}), \phi_1(\mathbf{x}), \dots, \phi_{M-1}(\mathbf{x}))^T$  represent our basis functions.

- The corresponding **conjugate prior** is given by a Gaussian distribution:

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

- As both the likelihood and the prior terms are Gaussians, the posterior distribution will also be Gaussian.
- If the posterior distributions  $p(\theta|\mathbf{x})$  are in the same family as the prior probability distribution  $p(\theta)$ , the prior and posterior are then called **conjugate distributions**, and the prior is called a **conjugate prior** for the likelihood.

Pause: why is the normal distribution's conjugate prior another normal?

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

$$p(\mu|\mathbf{X}) = \mathcal{N}(\mu|\mu_N, \sigma_N^2)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\mu_{\text{ML}}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

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

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

# Examples of conjugate priors

- Binomial:  $\beta$  prior
- Multinomial: Dirichlet prior
- Exponential, Poisson, or  $\gamma$ :  $\gamma$  prior
- Normal: Normal prior
- Uniform: Pareto prior

# Back to Bayesian Linear Regression

- Combining the **prior together with the likelihood** term:

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}, \mathbf{w}, \beta) \propto \left[ \prod_{n=1}^N \mathcal{N}(t_n | \mathbf{w}^T \boldsymbol{\phi}(\mathbf{x}_n), \beta^{-1}) \right] \mathcal{N}(\mathbf{w} | \mathbf{m}_0, \mathbf{S}_0).$$

- The **posterior** (with a bit of manipulation) takes the following Gaussian form:

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

where

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

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

- The posterior mean can be expressed in terms of the **least-squares estimator** and the **prior mean**:

$$\mathbf{m}_N = \mathbf{S}_N \left( \mathbf{S}_0^{-1} \mathbf{m}_0 + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \mathbf{w}_{ML} \right). \quad \boxed{\mathbf{w}_{ML} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t}.}$$

- As we increase our prior precision (decrease prior variance), we place greater weight on the prior mean relative to the data.

# Bayesian Linear Regression

- Consider a zero-mean, isotropic, Gaussian prior which is governed by a single precision parameter  $\alpha$ :

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

for which the posterior is Gaussian with:

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

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

- If we consider an infinitely broad prior,  $\alpha \rightarrow 0$ , the mean  $\mathbf{m}_N$  of the posterior distribution reduces to **maximum likelihood value**  $\mathbf{w}_{ML}$ . (*Can you see how?*)
- The log of the posterior distribution is given by the sum of the log-likelihood and the log of the prior:

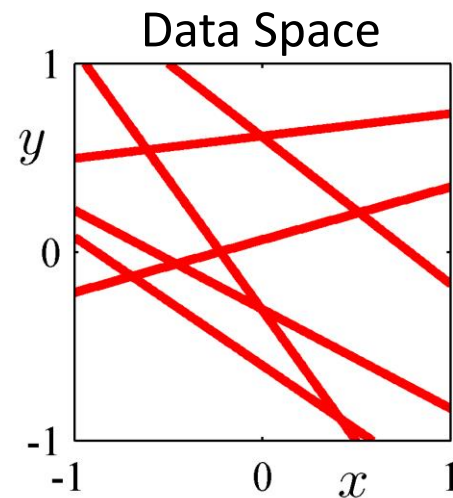
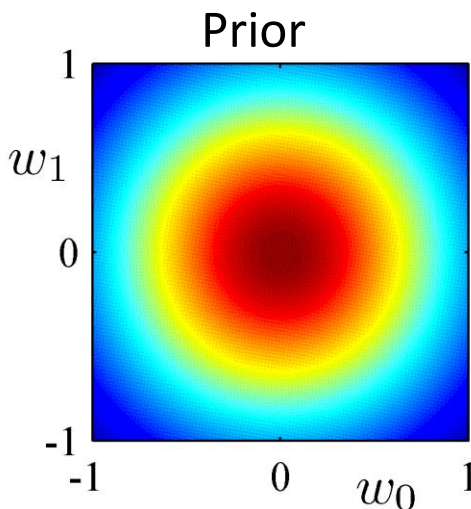
$$\ln p(\mathbf{w} | \mathcal{D}) = -\frac{\beta}{2} \sum_{n=1}^N (t_n - \mathbf{w}^T \phi(\mathbf{x}_n))^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} + \text{const.}$$

- Maximizing this posterior with respect to  $\mathbf{w}$  is equivalent to minimizing the **sum-of-squares error function** with a quadratic regulation term  $\lambda = \alpha / \beta$ .

# Bayesian Linear Regression

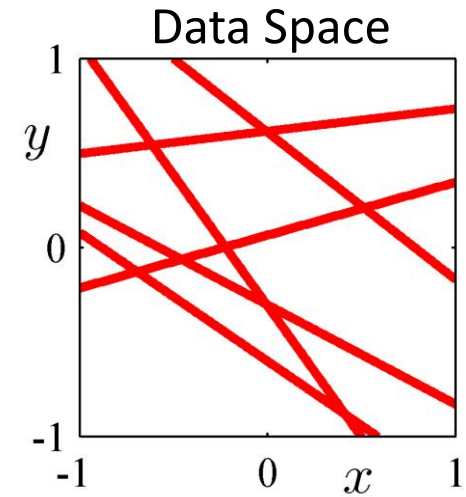
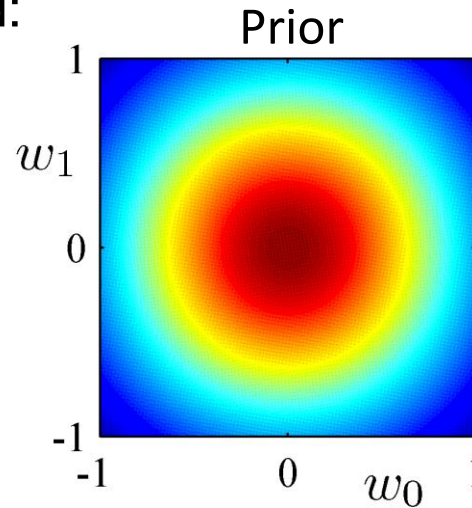
- Consider a linear model of the form:  $y(x, \mathbf{w}) = w_0 + w_1 x$ .
- The training data is generated from the function  $f(x, \mathbf{a}) = a_0 + a_1 x$  with  $a_0 = -0.3$  and  $a_1 = 0.5$  by first choosing  $x_n$  uniformly from  $[-1;1]$ , evaluating  $f(x, \mathbf{a})$ , and adding a small Gaussian noise.
- **Goal:** recover the values of  $a_0, a_1$  from such data.

When zero data points have been observed:

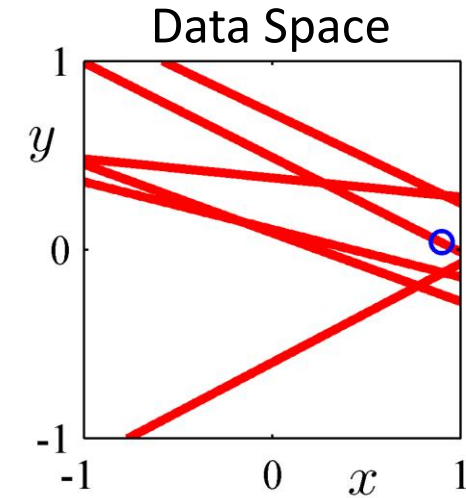
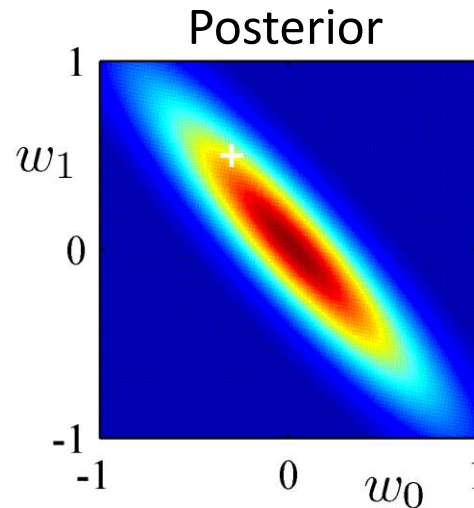
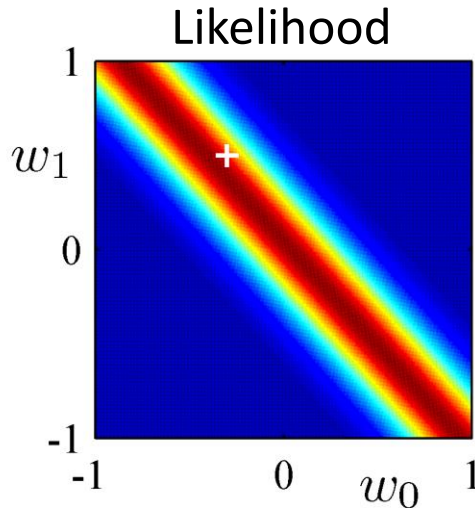


# Bayesian Linear Regression

0 data points are observed:



1 data point is observed:





# Bayesian Linear Regression

likelihood

prior/posterior

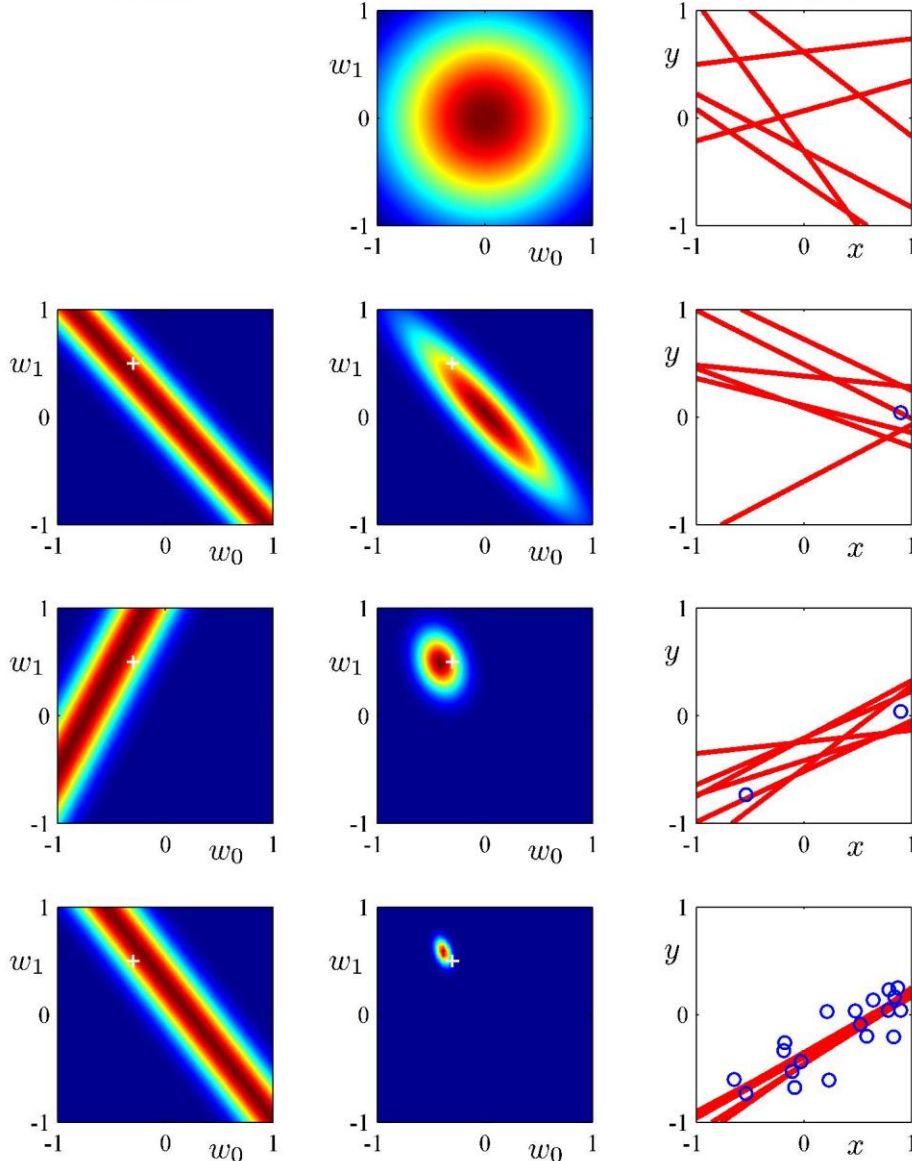
data space

0 data points are observed.

1 data point is observed.

2 data points are observed.

20 data points are observed.



# Predictive Distribution

- We can make predictions for a new input vector  $\mathbf{x}$  by **integrating over the posterior distribution**:

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


where

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

Noise in the  
target values



Uncertainty  
associated with  
parameter values.

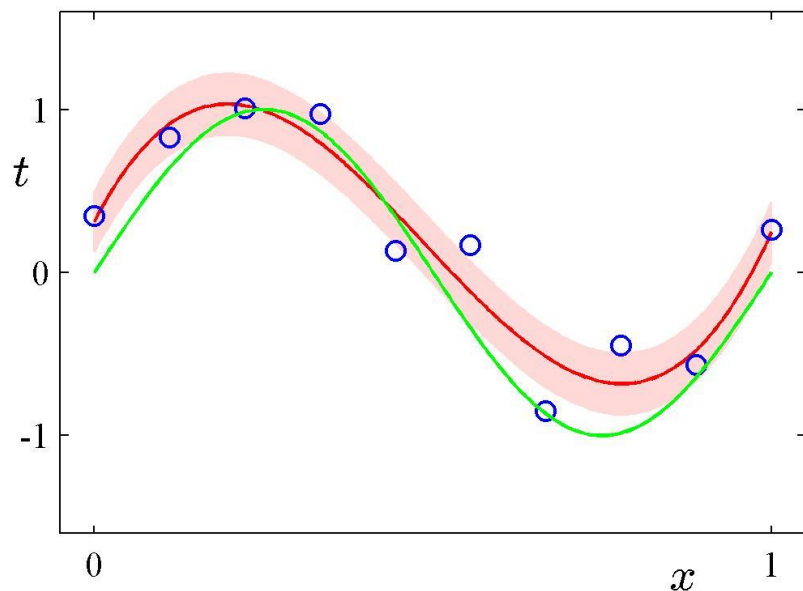


$\mathbf{m}_N$	$=$	$\beta \mathbf{S}_N \Phi^T \mathbf{t}$
$\mathbf{S}_N^{-1}$	$=$	$\alpha \mathbf{I} + \beta \Phi^T \Phi.$

- As  $N \rightarrow \infty$ :
  - The second term goes to zero
  - The variance of the predictive distribution arises only from the additive noise governed by parameter  $\beta$

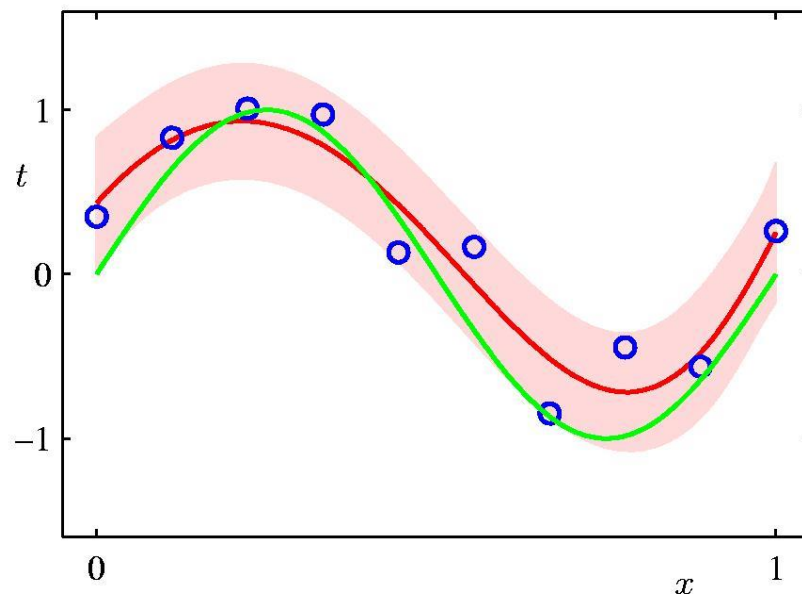
# Predictive Distribution: ML vs. Bayes

Predictive distribution based on maximum likelihood estimates



$$p(t|x, \mathbf{w}_{\text{ML}}, \beta_{\text{ML}}) = \mathcal{N}(t|y(x, \mathbf{w}_{\text{ML}}), \beta_{\text{ML}}^{-1})$$

Bayesian predictive distribution

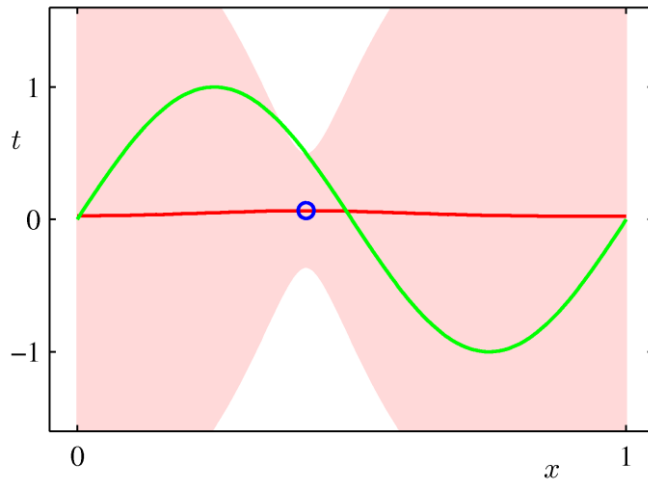


$$p(t|x, \mathbf{t}, \mathbf{X}) = \mathcal{N}(t|\mathbf{m}_N^T \phi(x), \sigma_N^2(x))$$

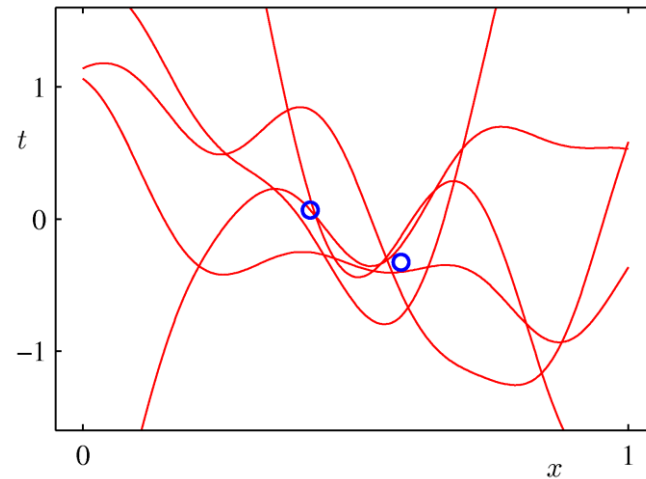
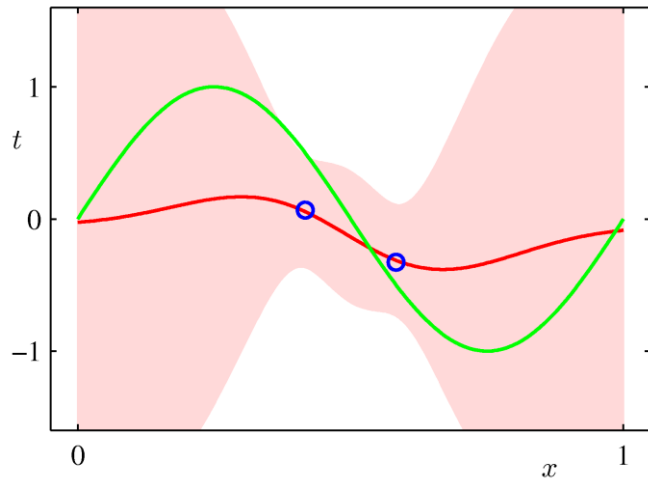
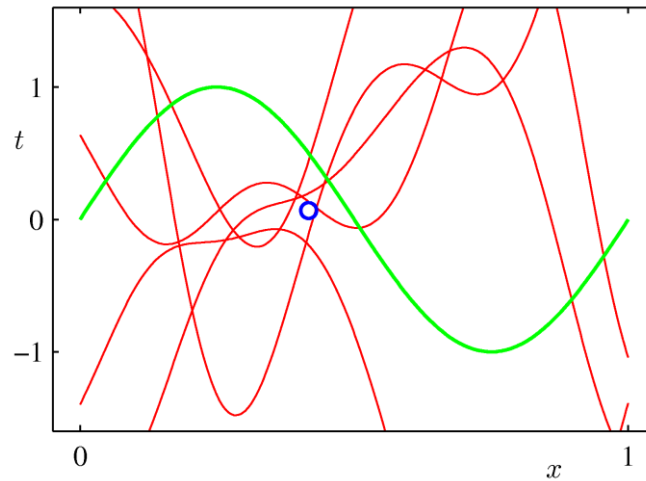
# Predictive Distribution

Sinusoidal dataset, nine Gaussian basis functions.

Predictive distribution



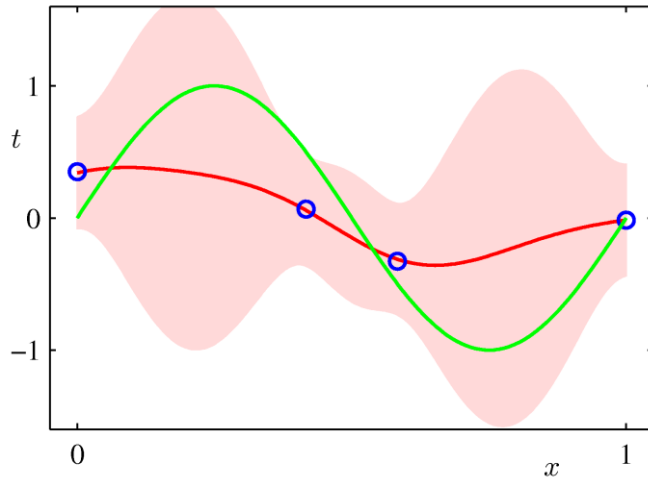
Samples from the posterior



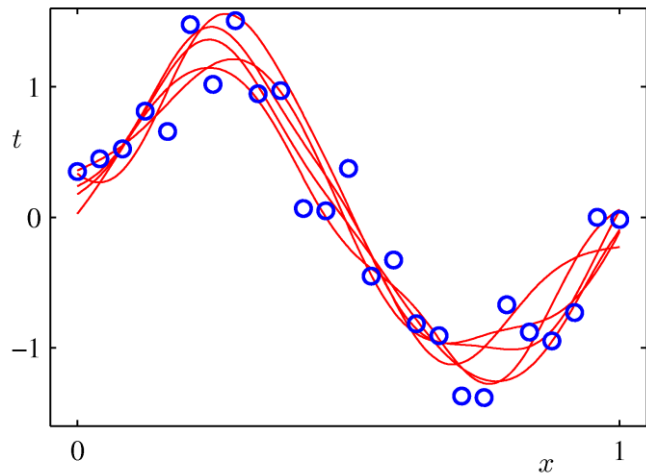
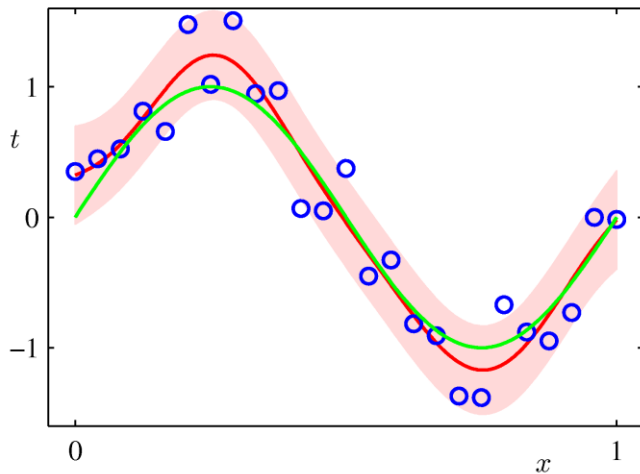
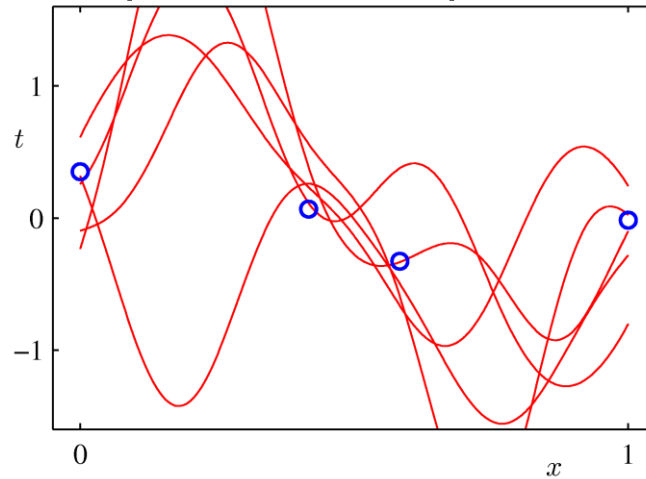
# Predictive Distribution

Sinusoidal dataset, nine Gaussian basis functions.

Predictive distribution



Samples from the posterior



# Bayesian Model Comparison

- The Bayesian view of model comparison involves the use of **probabilities to represent uncertainty** in the choice of the model.
- We would like to compare a set of  $L$  models  $\{\mathcal{M}_i\}$ , where  $i = 1, 2, \dots, L$ , using a training set  $\mathcal{D}$ .
- We **specify the prior distribution** over the different models  $p(\mathcal{M}_i)$ .
- Given a training set  $\mathcal{D}$ , we **evaluate the posterior**:

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

Posterior

Prior

*Model evidence or  
marginal likelihood*

- For simplicity, we will assume that all models are a-priori equally likely
- The model evidence expresses the preference shown by the data for different models.
- The ratio of two model evidences for two models is known as a **Bayes factor**:

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

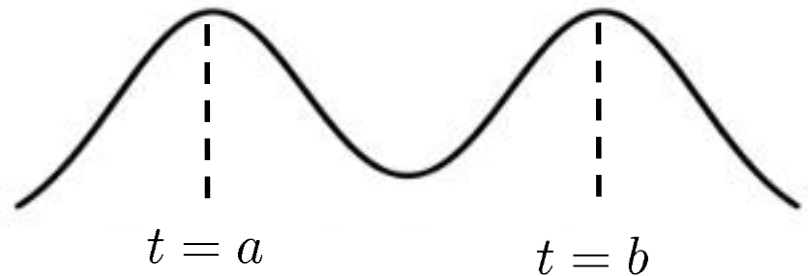
# Bayesian Model Comparison

- Once we compute the posterior  $p(M_i|\mathcal{D})$ , we can compute the **predictive (mixture) distribution**:

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

- The overall predictive distribution is obtained by **averaging the predictive distributions of individual models**, weighted by the posterior probabilities.

- For example, if we have two models, and one predicts a narrow distribution around  $t=a$  while the other predicts a narrow distribution around  $t=b$ , then the overall predictions will be bimodal:



- A simpler approximation, known as **model selection**, is to use the model with the highest evidence.

# Bayesian Model Comparison

- Remember, the posterior is given by

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

For a model governed by a set of parameters  $\mathbf{w}$ , the **model evidence** can be computed as follows:

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

- Observe that **the evidence is the normalizing term** that appears in the denominator in Bayes' rule:

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

- The model evidence is also often called **marginal likelihood**.



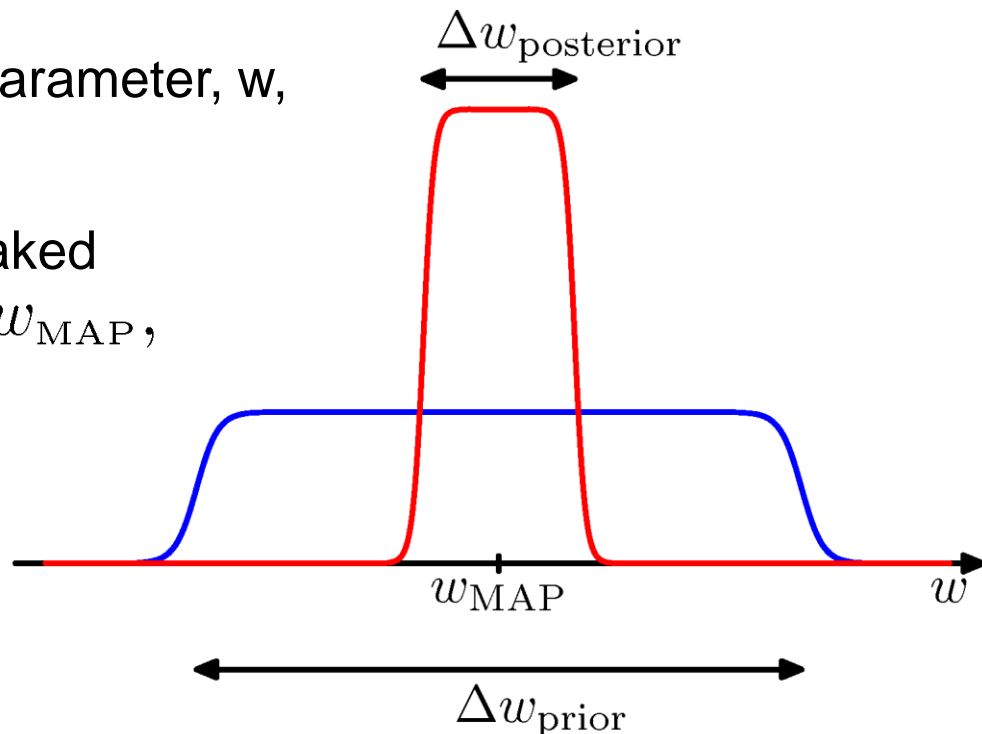
# Bayesian Model Comparison

- We next get some insight into the model evidence by making simple approximations.

- For a given model with a single parameter,  $w$ , consider approximations:

- Assume that **the posterior** is peaked around the most probable value  $w_{\text{MAP}}$ , with width  $\Delta w_{\text{posterior}}$

- Assume that **the prior** is flat with width  $\Delta w_{\text{prior}}$



$$p(\mathcal{D}) = \int p(\mathcal{D}|w)p(w) dw$$
$$\simeq p(\mathcal{D}|w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}$$

# Bayesian Model Comparison

- Taking the logarithms, we obtain:

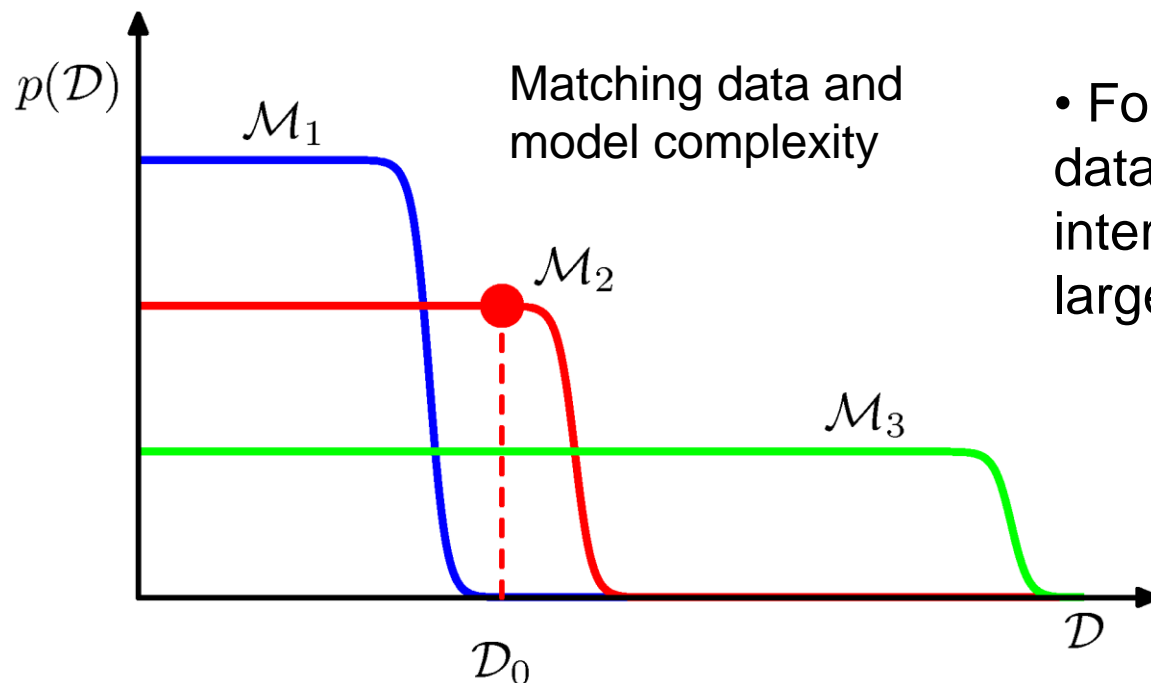
$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\text{MAP}}) + \underbrace{\ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)}_{\text{Negative}}.$$

- With  $M$  parameters, all assumed to have the same  $\Delta w_{\text{posterior}}/\Delta w_{\text{prior}}$  ratio:

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\mathbf{w}_{\text{MAP}}) + M \underbrace{\ln \left( \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right)}_{\text{Negative and linear in } M}.$$

- As we **increase the complexity** of the model (increase the number of adaptive parameters  $M$ ), **the first term will increase, whereas the second term will decrease** due to the dependence on  $M$ .
- The optimal model complexity: trade-off between these two competing terms.

# Bayesian Model Comparison



- For the particular observed dataset  $\mathcal{D}_0$ , the model  $\mathcal{M}_2$  with intermediate complexity has the largest evidence.

- The simple model cannot fit the data well, whereas the more complex model spreads its predictive probability and so assigns relatively small probability to any one of them.
- The marginal likelihood is **very sensitive to the prior used!**
- Computing the marginal likelihood makes sense only if you are certain about the choice of the prior.

# Hint for the question on slide 11

The exponent in the right-hand side is:

$$\begin{aligned} -\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2 - \frac{1}{2\sigma^2} \sum_{n=1}^N (x_n - \mu)^2 \\ = -\frac{\mu^2}{2} \left( \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2} \right) + \mu \left( \frac{\mu_0}{\sigma_0^2} + \frac{1}{\sigma^2} \sum_{n=1}^N x_n \right) + \text{const.} \end{aligned}$$

Compare this, term by term, to a single Gaussian's exponent:

$$-\frac{1}{2} \left( \frac{\mu - \mu_N}{\sigma_N} \right)^2 = -\frac{\mu^2}{2} \left( \frac{1}{\sigma_N^2} \right) + \mu \left( \frac{\mu_N}{\sigma_N^2} \right) + \text{const.}$$

And you arrive at the terms on the left-hand side of slide 12.