

# Machine Learning

Winter 2011/12

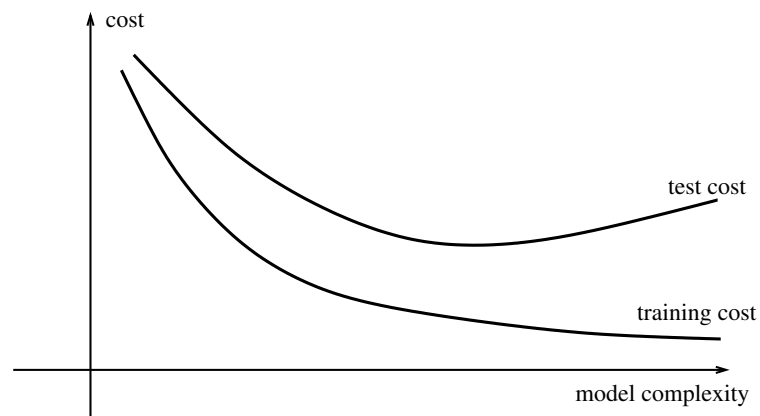
Roland Memisevic

Lecture 10, Jan. 16, 2012

## Outline

- ▶ Overfitting and ways to prevent it
- ▶ Bayesian reasoning
- ▶ Conjugate priors
- ▶ Evidence and model selection

## Overfitting and capacity control



## Views of overfitting

- ▶ Bias/Variance
- ▶ Size of the hypothesis space
- ▶ VC Dimension
- ▶ Curse of Dimensionality

## Preventing overfitting

- ▶ We have seen various approaches to preventing overfitting so far:
  1. Weight decay
  2. Early stopping
  3. Pseudo-counts (when estimating a discrete distribution)
  4. Choosing a constrained model class.
- ▶ These are also referred to as “**smoothing**” and “**regularization**”.
- ▶ We typically use *hyperparameters* to control the complexity (Eg., continuous  $\lambda$  for weight-decay, pseudo-count  $M$  to smooth a discrete distribution)
- ▶ There are various analytical approaches to choosing the right amount of smoothing: BIC, AIC, MDL, VC dimension.
- ▶ A more common approach:

## Inductive Bias

- ▶ Sometimes, we may have a principled way to reduce the size of the hypothesis class.
- ▶ *If you have any knowledge about the true model, use it!*
- ▶ Knowledge about the task can take many different forms, eg. we may know that
  - ▶ dependencies between inputs and outputs are linear; quadratic; sinusoidal, etc.,
  - ▶ classes are separated by large boundaries,
  - ▶ the data is structured like a sequence; tree; a grid, etc.
- ▶ Keep in mind that there can be *no learning without making assumptions* (“No Free Lunch”).
- ▶ In *Bayesian modeling* this observation is expressed through the requirement to specify a *prior distribution* before learning:

## Cross-validation

- ▶ Hold out **validation data** that is not used for training, and evaluate various hyperparameter settings on the validation data.
- ▶ A more common variation, that allows us to use all the data for training, is **cross-validation**:
  - ▶ Partition your training data into  $K$  groups, cycle through them, each time using  $K - 1$  of the subsets as training data, and one as validation data. This is known as  **$K$ -fold cross-validation**.
- ▶ The extreme case, where we use just one point at a time for validation, is known as leave-one-out cross-validation (LOOCV).

## Bayesian reasoning

- ▶ Practically all approaches to learning that we have discussed so far are based on *optimizing a cost function*.
- ▶ **Bayesian modeling** is a different way to learn from data, that is *not* based on optimization.
- ▶ It is based on letting both the data  $\mathcal{D}$  and the model parameters  $\theta$  be random variables.
- ▶ By using Bayes' rule we can then turn the task of learning into *probabilistic inference*:

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta) d\theta}$$

- ▶ Prior knowledge can be, and has to be, encoded in the form of a *prior distribution*  $p(\theta)$  over parameters.

## Bayesian reasoning

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta}) d\boldsymbol{\theta}}$$

- ▶ To get the **posterior** over parameters, multiply the **likelihood**  $p(\mathcal{D}|\boldsymbol{\theta})$  with the **prior**  $p(\boldsymbol{\theta})$ , and normalize.
- ▶ (This makes it necessary to interpret a probability as something other than a relative frequency, which has been the cause of many philosophical debates between “Bayesians”, who don’t mind adopting this interpretation, and “frequentists”, who do.)

## Predictive distribution

- ▶ The training data is typically a set of points  $\{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ .
- ▶ When treating parameters as random variables, *applying the model* to test data points  $\mathbf{x}$  amounts to using the rules of probability to compute the **predictive distribution**:

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}, \boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

- ▶ The same is true in conditional models (eg., regression, classification) of a target variable  $\mathbf{t}$  given  $\mathbf{x}$ , where the training data is given by pairs  $\{(\mathbf{x}_1, \mathbf{t}_1), \dots, (\mathbf{x}_N, \mathbf{t}_N)\}$ . The predictive distribution then takes the form

$$p(\mathbf{t}|\mathcal{D}, \mathbf{x}) = \int p(\mathbf{t}|\boldsymbol{\theta}, \mathbf{x})p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta}$$

## Bayesian reasoning

- ▶ Bayesian modeling provides us with a posterior distribution  $p(\boldsymbol{\theta}|\mathcal{D})$  over parameters *not* with a single point estimate for *optimal parameters*  $\boldsymbol{\theta}^*$ .
- ▶ In a setting where we get data sequentially, we can treat  $p(\boldsymbol{\theta}|\mathcal{D})$  as a prior and update it when more data arrives.
- ▶ This can be done the most conveniently, when we have a “conjugate” prior (see below).
- ▶ (We occasionally write  $p(\boldsymbol{\theta}|\mathcal{D})$  for both prior and posterior and let  $\mathcal{D}$  be the empty set when the prior is meant.)
- ▶ At the end of the day, we often still want a single answer, in which case we just need to use the rules of probability:

## Bayesian modeling

- ▶ The predictive distributions averages over many models, weighting them with their posterior.
- ▶ This tends to yield more accurate predictions than using a point estimate.
- ▶ The Bayesian modeling philosophy may be summarized as “Put everything that you know or assume on the table, then use the rules of probability to answer any question you may have.”
- ▶ In other words, we specify a joint distribution over all quantities of interest. The rest is probabilistic inference.
- ▶ The downside: Bayesian modeling can be computationally and mathematically demanding, and it often leads to intractable distributions that require approximate inference or sampling.

## Conjugate priors

- ▶ In some cases, we can find a combination of prior and likelihood that allows for closed form inference.
- ▶ This is usually the case when the product

$$p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})$$

has the same functional form as the constituents  $p(\mathcal{D}|\boldsymbol{\theta})$  and  $p(\boldsymbol{\theta})$ .

- ▶ This is known as **conjugacy**. And the prior, if it satisfies this property for a given likelihood function, is called a **conjugate prior**.
- ▶ Examples of conjugate priors include: the *beta distribution* (for a Bernoulli model), *Dirichlet distribution* (for a multinomial model), and the *normal-Wishart distribution* (for a Gaussian model).

## Bernoulli and beta distribution

- ▶ The conjugate prior is known as the

### beta distribution

$$\text{Beta}(\mu|a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \mu^{a-1} (1-\mu)^{b-1}$$

with mean and variance

$$\mathbb{E}[\mu] = \frac{a}{a+b}, \quad \text{var}[\mu] = \frac{ab}{(a+b)^2(a+b+1)}$$

where  $\Gamma(\cdot)$  in the normalizer is the Gamma-function:  
 $\Gamma(a) = \int_0^\infty u^{a-1} e^{-u} du$ , and  $a$  and  $b$  are positive, real-valued.

because

$$p(\mu|\mathcal{D}, a, b) \propto \mu^{m+a-1} (1-\mu)^{l+b-1}$$

## Bernoulli and beta distribution

- ▶ Recall that the **Bernoulli distribution** (“tossing a coin”) can be defined

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

where  $x$  is 0 or 1.

- ▶ The likelihood for a data-set  $\mathcal{D}$ , consisting of  $m$  1's and  $l = N - m$  0's, is

$$p(\mathcal{D}|\mu) = \prod_{n=1}^N \mu^{x_n} (1-\mu)^{1-x_n} = \mu^{\sum_n x_n} (1-\mu)^{\sum_n (1-x_n)} = \mu^m (1-\mu)^l$$

- ▶ In contrast to previously,  $\mu$  is now a random variable, so we write  $p(x|\mu)$  not  $p(x; \mu)$ .

## Predictive distribution and pseudo counts

- ▶  $p(\mu|\mathcal{D}, a, b)$  is just another beta distribution, and we can simply *read off* the normalizing constant:  $\frac{\Gamma(m+a+l+b)}{\Gamma(m+a)\Gamma(l+b)}$
- ▶ The **predictive distribution** for this model can be written

$$p(x=1|\mathcal{D}) = \int_0^1 p(x=1|\mu)p(\mu|\mathcal{D}) d\mu = \mu p(\mu|\mathcal{D}) d\mu$$

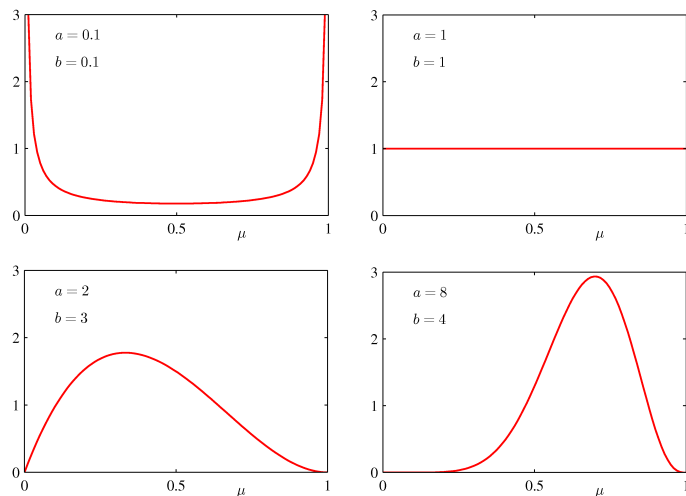
(so it is simply the expected value of the posterior)

- ▶ ...which after seeing the data ( $m$  1's and  $l$  0's) is

$$p(x=1|\mathcal{D}) = \frac{m+a}{m+a+l+b}$$

- ▶ This means that using the prior  $\text{Beta}(\mu|a, b)$  amounts to adding **pseudo counts**  $a$  and  $b$  to the maximum likelihood estimate  $\mu = \frac{m}{m+l}$  for  $\mu$ , lending justification to this way of smoothing the estimates!

## Beta distribution examples



## Multinomial and Dirichlet distribution

- ▶ The **discrete distribution** can be defined

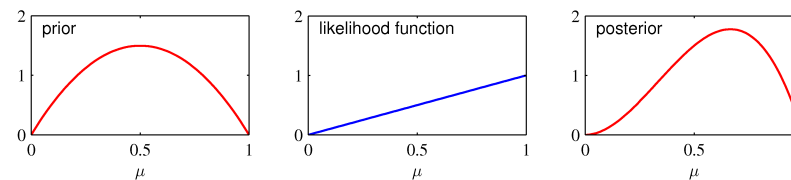
$$p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^K \mu_k^{x_k}$$

where  $\mathbf{x}$  is in one-hot encoding.

- ▶ (This is, obviously, a generalization of the Bernoulli)
- ▶ The likelihood for data  $\mathcal{D}$ , contained in some matrix  $\mathbf{X}$ , is

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^N \prod_{k=1}^K \mu_k^{x_{nk}} = \prod_{k=1}^K \mu_k^{\sum_n x_{nk}} =: \prod_{k=1}^K \mu_k^{m_k}$$

## Posterior $\propto$ likelihood $\times$ prior



- ▶ Note that both the likelihood and the prior are functions over  $\mu$  taking the same functional form.
- ▶ In this example, the prior is Beta(2, 2) and  $m = N = 1$ ,  $l = N - m = 0$ , the posterior is Beta(3, 2)
- ▶ Note that *few observations* change the prior only slightly.
- ▶ For *many observations* we would get something closer to the maximum-likelihood estimate.

## Multinomial and Dirichlet distribution

- ▶ The conjugate prior is the

### Dirichlet distribution

$$\text{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1) \cdots \Gamma(\alpha_K)} \prod_{k=1}^K \mu_k^{\alpha_k - 1}, \quad \alpha_0 = \sum_k \alpha_k$$

with mean and variance

$$\mathbb{E}[\mu_k] = \frac{\alpha_k}{\alpha_0}, \quad \text{var}[\mu_k] = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)}$$

where  $\alpha_k$  are positive, real-valued.

because

$$p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha}) \propto \prod_{k=1}^K \mu_k^{\alpha_k + m_k - 1}$$

## Multinomial and Dirichlet distribution

- ▶  $p(\boldsymbol{\mu}|\mathcal{D}, \boldsymbol{\alpha})$  is just another Dirichlet distribution, and so we can, again, read off the normalizing constant:

$$\frac{\Gamma(\alpha_0 + N)}{\Gamma(\alpha_1 + m_1), \dots, \Gamma(\alpha_K + m_K)}$$

- ▶ The means of the posterior (and, as it turns out, parameters of the predictive distribution) are again given by adding **pseudo counts** to the maximum likelihood estimates:

$$\mathbb{E}[\mu_k] = \frac{m_k + \alpha_k}{\alpha_0}$$

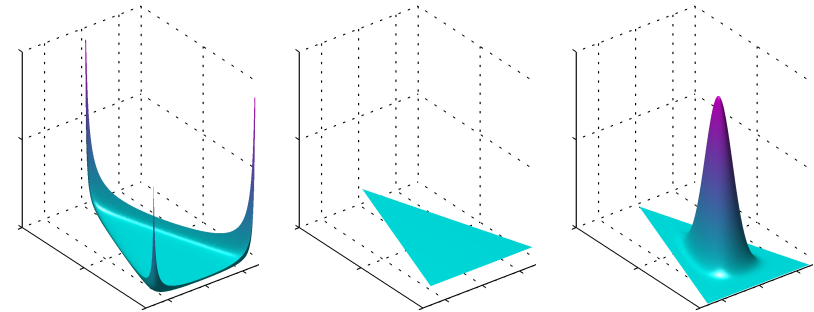
again lending justification to the common procedure of parameter smoothing.

- ▶ Note again, how this will make most of a difference for small data sets.

## Conjugates for the 1-D Gaussian

- ▶ Since the likelihood function for a 1-dimensional Gaussian with *unknown* mean  $\mu$  and *known* precision (inverse covariance)  $\lambda$  is the exponential of a quadratic function of  $\mu$ , the conjugate prior over  $\mu$  is a Gaussian  $\mathcal{N}(\mu|\mu_0, \sigma_0^2)$ .
- ▶ The posterior turns out to be a Gaussian with mean  $\frac{\sigma^2}{N\sigma_0^2 + \sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2 + \sigma^2}\frac{\sum_n x_n}{N}$  and precision  $\frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$ .
- ▶ For a one-dimensional Gaussian with *known* mean  $\mu$  and *unknown* precision  $\lambda$ , the conjugate prior over  $\lambda$  is the Gamma distribution  $\text{Gam}(\lambda|a, b) = \frac{1}{\Gamma(a)}b^a\lambda^{a-1}\exp(-b\lambda)$
- ▶ For a one-dimensional Gaussian with *unknown* mean  $\mu$  and *unknown* precision  $\lambda$ , the conjugate prior over  $\mu$  and  $\lambda$  is the product  $p(\mu, \lambda) = \mathcal{N}(\mu|\mu_0, (\beta\lambda)^{-1})\text{Gam}(\lambda|a, b)$

## Dirichlet distribution examples



- ▶ Dirichlet distributions with all  $\alpha_1 = \dots = \alpha_K = \{0.1(\text{left}), 1(\text{middle}), 10(\text{right})\}$
- ▶ The Dirichlet distribution is confined to the *simplex*.

## Conjugates for the multivariate Gaussian

- ▶ For a  $D$ -dimensional Gaussian with *unknown* mean  $\boldsymbol{\mu}$  and *known* precision matrix  $\boldsymbol{\Lambda}$ , the conjugate prior over  $\boldsymbol{\mu}$  is a multivariate Gaussian (not surprisingly).
- ▶ For a  $D$ -dimensional Gaussian with *known* mean  $\boldsymbol{\mu}$  and *unknown* precision matrix  $\boldsymbol{\Lambda}$ , the conjugate prior over  $\boldsymbol{\Lambda}$  is the *Wishart distribution*.
- ▶ For a  $D$ -dimensional Gaussian with *unknown* mean  $\boldsymbol{\mu}$  and *unknown* precision  $\boldsymbol{\Lambda}$ , the conjugate prior over  $\boldsymbol{\mu}$  and  $\boldsymbol{\Lambda}$  is the product of a multivariate Gaussian and a Wishart distribution.

## Bayesian linear regression

- ▶ Recall that the linear regression model can be defined as the conditional Gaussian

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

which, again, is the exponential of a quadratic function of  $\mathbf{w}$ .

- ▶ The conjugate prior is therefore also a Gaussian

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

- ▶ To show this, complete the square in the exponential, or use Gaussian identities (eg., Bishop page 93).

## MAP estimate and regularization

- ▶ The maximum of the posterior is known as the **maximum a-posteriori (MAP)** estimate.
- ▶ The MAP estimate of any regression model with a zero-mean Gaussian prior with spherical covariance matrix is the same as regression with weight-decay penalty on the parameters:

$$\begin{aligned} \log p(\mathbf{w}|\mathcal{D}) &= \text{const.} + \log p(\mathcal{D}|\mathbf{w}) + \log p(\mathbf{w}) \\ &= \text{const.} - \frac{\beta}{2} \sum_{n=1}^N (t_n - f(\mathbf{x}_n; \mathbf{w}))^2 - \frac{\alpha}{2} \mathbf{w}^T \mathbf{w} \end{aligned}$$

- ▶ So it is not surprising that we get the ridge-regression solution as the mean/mode of the Gaussian in the previous example.

## Bayesian linear regression

- ▶ A common choice in practice is

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

for some precision parameter  $\alpha$ .

- ▶ After seeing data  $\mathcal{D} = \{(\mathbf{x}_1, t_1), \dots, (\mathbf{x}_N, t_N)\}$ , which we can stack in matrices  $\mathbf{X}$  and  $\mathbf{t}$ , we get the posterior

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

with

$$\mathbf{m}_N = \beta \mathbf{S}_N \mathbf{X}^T \mathbf{t}$$

and

$$\mathbf{S}_N^{-1} = \alpha \mathbf{I} + \beta \mathbf{X}^T \mathbf{X}$$

## Predictive distribution

- ▶ The predictive distribution is given by

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

- ▶ This can be simplified to

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

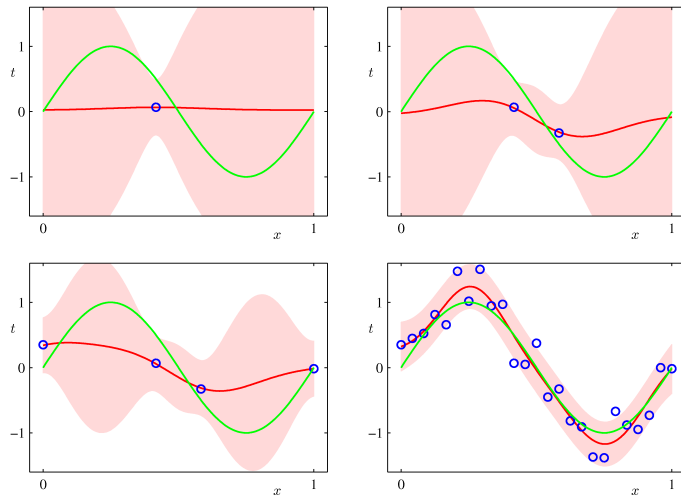
where

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

so the variance depends on  $\mathbf{x}$ .

- ▶ The mean of the predictive distribution is the same as the predictions made with the weight-decay regularized regression model.
- ▶ Since we have a full distribution over parameters, we also have input-dependent error-bars:

## Predictive distribution example



## Evidence and Model Selection

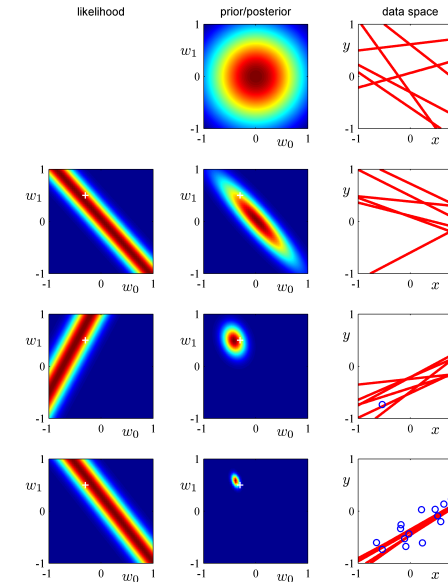
- ▶ The Bayes rule normalizing constant

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

is also known as **evidence**.

- ▶ It integrates out the parameters to represent *the probability of the data set, given the type of model*.
- ▶ Being able to compute  $p(\mathcal{D})$  allows us to perform **model selection** without cross-validation:

## Bayesian 1-D regression example revisited



## Evidence and Model Selection

- ▶ Assume you have multiple types of model  $\mathcal{M}_1, \dots, \mathcal{M}_L$  (for example, polynomial regression of different orders; a set of different classifiers; number of cluster centers in a mixture model; etc.).
- ▶ To perform model selection, define a prior  $p(\mathcal{M}_i)$  over model classes and compute the posteriors

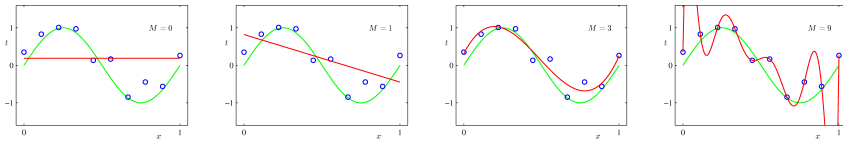
$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{D}|\mathcal{M}_i)p(\mathcal{M}_i)$$

where  $p(\mathcal{D}|\mathcal{M}_i)$  is the evidence for model type  $\mathcal{M}_i$ .

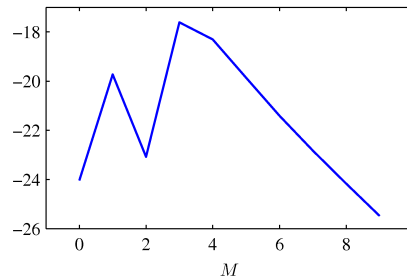
- ▶ This allows us to pick the right model given the data.
- ▶ Alternatively, we can make predictions by using the rules of probability and averaging (marginalize) over models, which amounts to weighting each model with  $p(\mathcal{M}_i|\mathcal{D})$ .
- ▶ For Gaussians/linear regression, it is possible to compute  $p(\mathcal{D})$  in closed form (see Bishop, page 169).

## Polynomial regression revisited

The polynomial regression example from Lecture 2:



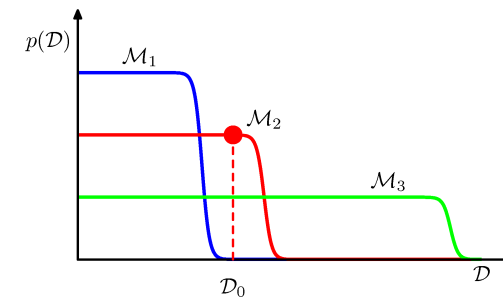
The log-evidence as a function of the order,  $M$ :



## Non-conjugate priors

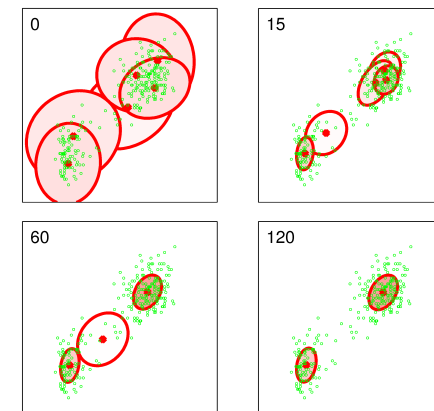
- ▶ A fully Bayesian treatment is possible in closed form only for the simplest kinds of model.
- ▶ Bayesian modeling is therefore one of the main applications of *approximate inference and sampling*.
- ▶ Some examples from the Bishop book:
  - ▶ Logistic regression (Laplace approximation), page 217
  - ▶ Logistic regression (Variational inference), page 498
  - ▶ Backprop networks (Laplace approximation), page 277
  - ▶ Mixture of Gaussians (Variational inference), page 474

## Evidence intuition



- ▶ Simple models assign a lot of probability mass to a small number of data-sets.
- ▶ Complex models spread their probability mass over many data-sets.
- ▶ Evidence can help choose the right model complexity.

## A Bayesian mixture model can infer the number of clusters from data



- ▶ Training iteration shown in the top-left corners.
- ▶ Gaussians with vanishing mixing proportions not shown.