Linear Models for Regression

Seungjin Choi

Department of Computer Science and Engineering
Pohang University of Science and Technology
77 Cheongam-ro, Nam-gu, Pohang 37673, Korea
seungjin@postech.ac.kr

Outline

- Regression
- Least Squares Method
- Ridge Regression
- Least Mean Squares
- Recursive (Sequential) least squares
- Bias-Variance Dilemman
- Bayesian Linear Regression

Regression

- Regression aims at modeling the dependence of a response Y on a covariate X. In other words, the goal of regression is to predict the value of one or more continuous target variables y given the value of input vector $\boldsymbol{x}$.
- The regression model is described by

$$y = f(\boldsymbol{x}) + \epsilon.$$

- Terminology:
  - $\boldsymbol{x}$: input, independent variable, predictor, regressor, covariate
  - y: output, dependent variable, response
- The dependence of a response on a covariate is captured via a conditional probability distribution, $p(y|\boldsymbol{x})$.
- Depending on $f(\boldsymbol{x})$,
  - Linear regression: $f(\boldsymbol{x}) = \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}) + w_0$.
  - Kernel regression: $f(\boldsymbol{x}) = \sum_{i=1}^{N} w_i k(\boldsymbol{x}, \boldsymbol{x}_i) + w_0$.

Regression Function: Conditional Mean

We consider the mean squared error and find the MMSE estiamte:

$$\begin{aligned}
\mathcal{E}(f) &= \left\langle \|y - f(x)\|^2 \right\rangle \\
&= \int \int \|y - f(x)\|^2 p(x, y) dx dy \\
&= \int \int \|y - f(x)\|^2 p(x) p(y|x) dx dy \\
&= \int p(x) \left[\underbrace{\int \|y - f(x)\|^2 p(y|x) dy}_{\text{to be minimized}} \right] dx
\end{aligned}$$

$\frac{\partial}{\partial f(x)} \left[\int \|y - f(x)\|^2 p(y|x) dy \right] = 0 \Rightarrow \boxed{f(x) = \int y p(y|x) dy = \langle y|x \rangle}$.

Linear Regression

Linear regression refers to a model in which the conditional mean of y given the value of $\boldsymbol{x}$ is an affine function of $\phi(\boldsymbol{x})$

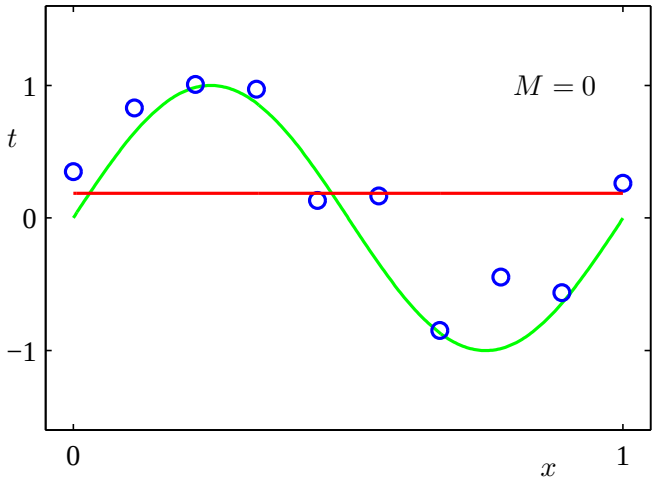
$$f(\boldsymbol{x}) = \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}) + w_0 \phi_0(\boldsymbol{x}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}),$$

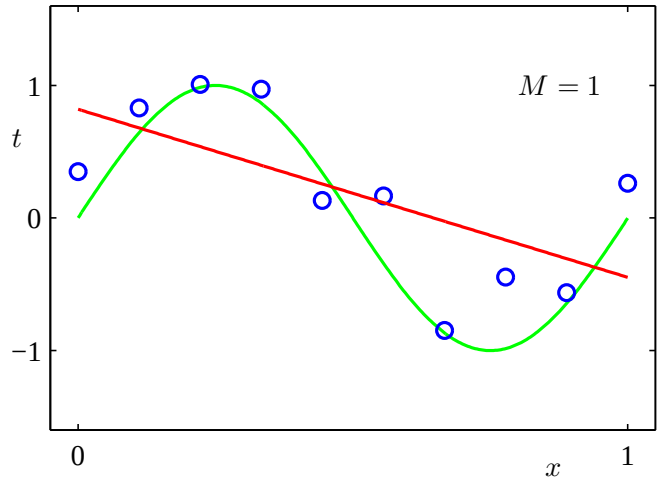
where $\phi_j(\boldsymbol{x})$ are known as basis functions and

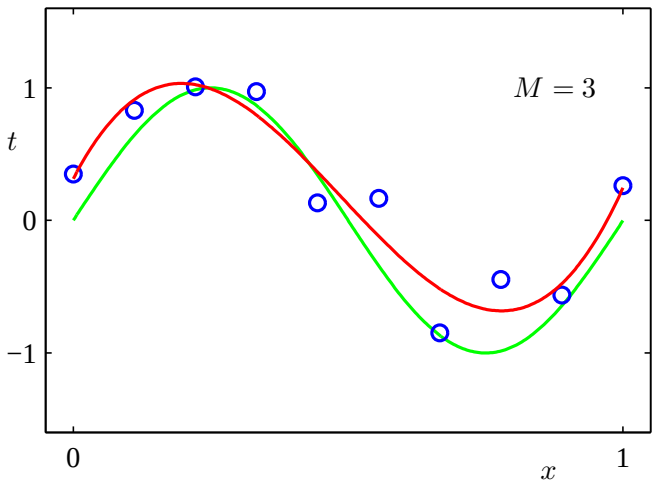
$$\begin{aligned} \boldsymbol{w} &= [w_0, w_1, \ldots, w_M]^\top, \\ \boldsymbol{\phi} &= [\phi_0, \phi_1, \ldots, \phi_M]^\top. \end{aligned}$$

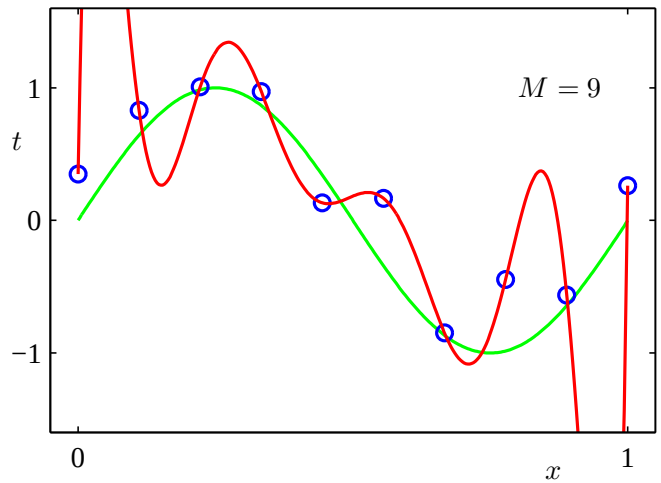
By using nonlinear basis functions, we allow the function $f(\boldsymbol{x})$ to be a nonlinear function of the input vector $\boldsymbol{x}$ (but a linear function of $\phi(\boldsymbol{x})$).

Polynomial Regression: $y_t = \sum_{j=0}^{M} w_j \phi_j(x_t) = \sum_{j=0}^{M} w_j x_t^j$









Basis Functions

- Polynomial regression: $\phi_j(x) = x^j$.
- Gaussian basis functions: $\phi_j(\boldsymbol{x}) = \exp\left\{-\frac{\|\boldsymbol{x}-\boldsymbol{\mu}_j\|^2}{2\sigma^2}\right\}$.
- Spline basis functions: Piecewise polynomials (divide the input space up into regions and fit a different polynomial in each region).
- Many other possible basis functions: sigmoidal basis functions, hyperbolic tangent basis functions, Fourier basis, wavelet basis, and so on.

Least Squares Method

Given a set of training data $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^N$, we determine the weight vector $\boldsymbol{w}$ which minimizes

$$\mathcal{E}_{LS} = \frac{1}{2}\sum_{t=1}^{N}\left\{y_t - \boldsymbol{w}^\top \phi(\boldsymbol{x}_t)\right\}^2 = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}\|^2,$$

where $\boldsymbol{y} = [y_1, \ldots, y_N]^\top$ and $\boldsymbol{\Phi} \in \mathbb{R}^{N\times(M+1)}$ is known as the design matrix with $\Phi_{tj} = \phi_j(\boldsymbol{x}_t)$, i.e.,

$$\boldsymbol{\Phi} = \begin{bmatrix} \phi_0(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_1) & \cdots & \phi_M(\boldsymbol{x}_1) \\ \phi_0(\boldsymbol{x}_2) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_M(\boldsymbol{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_0(\boldsymbol{x}_N) & \phi_1(\boldsymbol{x}_N) & \cdots & \phi_M(\boldsymbol{x}_N) \end{bmatrix}.$$

$\frac{\partial \mathcal{E}_{LS}}{\partial \boldsymbol{w}} = 0$ leads to the normal equation that is of the form

$$\boldsymbol{\Phi}^\top \boldsymbol{\Phi} \boldsymbol{w} = \boldsymbol{\Phi}^\top \boldsymbol{y}.$$

Thus, LS estimate of $\boldsymbol{w}$ is given by

$$\boldsymbol{w}_{LS} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y} = \boldsymbol{\Phi}^\dagger \boldsymbol{y},$$

where $\boldsymbol{\Phi}^\dagger$ is known as the Moore-Penrose pseudo-inverse.
$\boldsymbol{\Phi}^\top \boldsymbol{\Phi}$ is known as the Gram matrix.

Maximum Likelihood

We assume that the target variable y_t is given by a deterministic function $f(\boldsymbol{x}_t, \boldsymbol{w}) = \boldsymbol{w}^\top \phi(\boldsymbol{x}_t)$ with additive Gaussian noise so that

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{w} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$.
The log-likelihood is given by

$$\begin{aligned} \mathcal{L} &= \log p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}) = \sum_{t=1}^{N} \log p(y_t|\phi(\boldsymbol{x}_t), \boldsymbol{w}) \\ &= -\frac{N}{2}\log\sigma^2 - \frac{N}{2}\log 2\pi - \sigma^{-2}\mathcal{E}_{LS}. \end{aligned}$$

Therefore, under Gaussian noise assumption, $\boxed{\boldsymbol{w}_{ML} = \boldsymbol{w}_{LS}}$.

Ridge Regression

We consider the sum-of-squares error function with a Euclidean norm-based regularizer

$$\mathcal{E} = \underbrace{\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{w}\|^2}_{\textit{LSfit}} + \underbrace{\frac{\lambda}{2}\|\boldsymbol{w}\|^2}_{\textit{regularizer}}.$$

Solving $\frac{\partial \mathcal{E}}{\partial \boldsymbol{w}} = 0$ for $\boldsymbol{w}$ leads to

$$\boldsymbol{w}_{\textit{ridge}} = \left(\lambda \boldsymbol{I} + \boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{y}.$$

Ridge Regression: MAP Perpective

Recall the likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w}) = \mathcal{N}(\boldsymbol{y}|\boldsymbol{\Phi}\boldsymbol{w}, \sigma^2\boldsymbol{I}).$$

Assume a zero-mean Gaussian prior with covariance $\boldsymbol{\Sigma}$ for parameters $\boldsymbol{w}$:

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|0, \boldsymbol{\Sigma}).$$

Then the posterior over $\boldsymbol{w}$,

$$p(\boldsymbol{w}|\boldsymbol{y}, \boldsymbol{\Phi}) = \frac{p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w})p(\boldsymbol{w})}{\int p(\boldsymbol{y}|\boldsymbol{\Phi}, \boldsymbol{w})p(\boldsymbol{w})d\boldsymbol{w}},$$

is still Gaussian with mean and mode at

$$\widehat{\boldsymbol{w}} = (\sigma^2\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y}.$$

When $\boldsymbol{\Sigma}$ is proportional to identity, i.e., $\boldsymbol{\Sigma} = \lambda^{-1}\sigma^2\boldsymbol{I}$, this is called ridge regression.

Ridge Regression: Shrinkage

- Suppose that the SVD of the design matrix $\boldsymbol{\Phi} \in \mathbb{R}^{N\times(M+1)}$ has the form

$$\boldsymbol{\Phi} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top, \quad \text{where } \boldsymbol{D} = \text{diag}(d_1, \ldots, d_{M+1}).$$

- Using the SVD we write the least squares fitted vector as

$$\boldsymbol{\Phi}\boldsymbol{w}_{LS} = \boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} = \boldsymbol{U}\boldsymbol{U}^\top\boldsymbol{y} = \sum_{i=1}^{M+1}\boldsymbol{u}_i\boldsymbol{u}_i^\top\boldsymbol{y},$$

  where $\boldsymbol{U}^\top\boldsymbol{y}$ are the coordinates of $\boldsymbol{y}$ with respect to the orthonormal basis $\boldsymbol{U}$.

- The ridge solutions are

$$\begin{aligned}\boldsymbol{\Phi}\boldsymbol{w}_{ridge} &= \boldsymbol{\Phi}\left(\lambda\boldsymbol{I} + \boldsymbol{\Phi}^\top\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\boldsymbol{y} \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^2 + \lambda\boldsymbol{I})^{-1}\boldsymbol{D}\boldsymbol{U}^\top\boldsymbol{y} = \sum_{i=1}^{M+1}\boldsymbol{u}_i\frac{d_i^2}{d_i^2+\lambda}\boldsymbol{u}_i^\top\boldsymbol{y}.\end{aligned}$$

  It shrinks the coordinates by the factors $\frac{d_i^2}{d_i^2+\lambda}$. A greater amount of shrinkage is applied to basis vectors with smaller d_i^2.

Least Mean Square (LMS)

LMS is a gradient-descent method which minimizes the instantaneous error $\mathcal{E}_t$, where

$$\mathcal{E}_{LS} = \sum_{t=1}^{N} \mathcal{E}_t = \frac{1}{2} \sum_{t=1}^{N} \left\{ y_t - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_t) \right\}^2 .$$

The gradient descent method leads to the updating rule for $\boldsymbol{w}$ that is of the form

$$\begin{aligned} \boldsymbol{w} &\leftarrow \boldsymbol{w} - \eta \nabla \mathcal{E}_t \\ &\leftarrow \boldsymbol{w} + \eta \left\{ y_t - \boldsymbol{w}^\top \boldsymbol{\phi}(\boldsymbol{x}_t) \right\} \boldsymbol{\phi}(\boldsymbol{x}_t), \end{aligned}$$

where $\eta > 0$ is a constant known as learning rate.

Recursive (Sequential) LS

We introduce the forgetting factor λ to de-emphasize old samples, leading to the following error function

$$\mathcal{E}_{RLS} = \frac{1}{2}\sum_{i=1}^{t} \lambda^{t-i}(y_i - \boldsymbol{\phi}_i \boldsymbol{w}_t^\top)^2,$$

where $\boldsymbol{\phi}_t = \boldsymbol{\phi}(\boldsymbol{x}_t)$.
Solving $\frac{\partial \mathcal{E}_{RLS}}{\partial \boldsymbol{w}_t} = 0$ for $\boldsymbol{w}_t$ leads to

$$\left[\sum_{i=1}^{t} \lambda^{t-i}\boldsymbol{\phi}_i\boldsymbol{\phi}_i^\top\right] \boldsymbol{w}_t = \left[\sum_{i=1}^{t} \lambda^{t-i} y_i \boldsymbol{\phi}_i\right].$$

We define

$$\begin{aligned} \boldsymbol{P}_t &= \left[\sum_{i=1}^{t} \lambda^{t-i} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top\right]^{-1}, \\ \boldsymbol{r}_t &= \left[\sum_{i=1}^{t} \lambda^{t-i} y_i \boldsymbol{\phi}_i\right]. \end{aligned}$$

With these definitions, we have

$$\boldsymbol{w}_t = \boldsymbol{P}_t \boldsymbol{r}_t.$$

The core idea of RLS is to apply the matrix inversion lemma to develop the sequential algorithm without matrix inversion.

$$(\boldsymbol{A} + \boldsymbol{BCD})^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{B}(\boldsymbol{DA}^{-1}\boldsymbol{B} + \boldsymbol{C}^{-1})^{-1}\boldsymbol{DA}^{-1}.$$

The recursion for $\boldsymbol{P}_t$ is given by

$$\begin{aligned}
\boldsymbol{P}_t &= \left[\sum_{i=1}^{t} \lambda^{t-i} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top\right]^{-1} \\
&= \left[\lambda \sum_{i=1}^{t-1} \lambda^{t-1-i} \boldsymbol{\phi}_i \boldsymbol{\phi}_i^\top + \boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top\right]^{-1} \\
&= \frac{1}{\lambda}\left[\boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1}\boldsymbol{\phi}_t \boldsymbol{\phi}_t^\top \boldsymbol{P}_{t-1}}{\lambda + \boldsymbol{\phi}_t^\top \boldsymbol{P}_{t-1}\boldsymbol{\phi}_t}\right]. \quad \text{(matrix inversion lemma)}
\end{aligned}$$

Thus, the updating rule for $\boldsymbol{w}$ is given by

$$\begin{aligned}
\boldsymbol{w}_t &= \boldsymbol{P}_t \boldsymbol{r}_t \\
&= \frac{1}{\lambda}\left[\boldsymbol{P}_{t-1} - \frac{\boldsymbol{P}_{t-1}\phi_t\phi_t^\top \boldsymbol{P}_{t-1}}{\lambda + \phi_t^\top \boldsymbol{P}_{t-1}\phi_t}\right][\lambda \boldsymbol{r}_{t-1} + y_t\phi_t] \\
&= \boldsymbol{w}_{t-1} + \underbrace{\frac{\boldsymbol{P}_{t-1}\phi_t}{\lambda + \phi_t^\top \boldsymbol{P}_{t-1}\phi_t}}_{\textit{gain}} \underbrace{\left[y_t - \phi_t^\top \boldsymbol{w}_{t-1}\right]}_{\textit{error}}.
\end{aligned}$$

Expected Loss

The squared loss L is given by

$$\boxed{L(y, f(\boldsymbol{x})) = (f(\boldsymbol{x}) - y)^2 .}$$

The expected loss is computed by

$$\begin{aligned}
\mathbb{E}\{L(y, f(\boldsymbol{x}))\} &= \int\int (f(\boldsymbol{x}) - y)^2 \, p(\boldsymbol{x}, y) d\boldsymbol{x} dy \\
&= \mathbb{E}\left\{(f(\boldsymbol{x}) - \mathbb{E}\{y|\boldsymbol{x}\} + \mathbb{E}\{y|\boldsymbol{x}\} - y)^2\right\} \\
&= \mathbb{E}\left\{(f(\boldsymbol{x}) - \mathbb{E}\{y|\boldsymbol{x}\})^2 + (\mathbb{E}\{y|\boldsymbol{x}\} - y)^2\right\} \\
&\quad + \underbrace{2\mathbb{E}\left\{(f(\boldsymbol{x}) - \mathbb{E}\{y|\boldsymbol{x}\})(\mathbb{E}\{y|\boldsymbol{x}\} - y)\right\}}_{0} .
\end{aligned}$$

The expected loss can be written as

$$\mathbb{E}\{L(y, f(\boldsymbol{x}))\} = \underbrace{\int (f(\boldsymbol{x}) - \mathbb{E}\{y|\boldsymbol{x}\})^2 \, p(\boldsymbol{x}) d\boldsymbol{x}}_{\text{first term}} + \underbrace{\int (\mathbb{E}\{y|\boldsymbol{x}\} - y)^2 \, p(\boldsymbol{x}) d\boldsymbol{x}}_{\text{second term}}.$$

- The function $f(\boldsymbol{x})$ appears only in the first term which will be minimized when $f(\boldsymbol{x}) = \mathbb{E}\{y|\boldsymbol{x}\}$.
- The second term is the variance of the distribution of y, averaged over $\boldsymbol{x}$, representing the intrinsic variability of the target data and can be regarded as noise. Because it is independent of $f(\boldsymbol{x})$, it represents the irreducible minimum value of the loss function.

Bias-Variance Decomposition

Consider the integrand of the first term, which for a particular data set $\mathcal{D}$ takes the form $(f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}\{y|\boldsymbol{x}\})^2$.
Then we have

$$\mathbb{E}_{\mathcal{D}}\left\{(f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}\{y|\boldsymbol{x}\})^2\right\} = \underbrace{(\mathbb{E}_{\mathcal{D}}\{f(\boldsymbol{x};\mathcal{D})\} - \mathbb{E}\{y|\boldsymbol{x}\})^2}_{(\text{bias})^2} + \underbrace{\mathbb{E}_{\mathcal{D}}\left\{(f(\boldsymbol{x};\mathcal{D}) - \mathbb{E}_{\mathcal{D}}\{f(\boldsymbol{x};\mathcal{D})\})^2\right\}}_{\text{variance}}.$$

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise}.$$

Bias-Variance Dilemma

There is a trade-off between bias and variance:

- flexible models: low bias but high variance
- rigid models: high bias but low variance

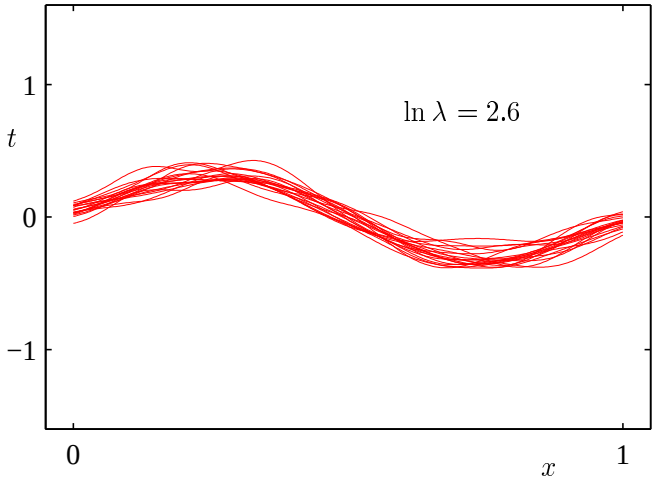
The model with the optimal predictive capability is the one that leads to the best balance between bias and variance.

Example

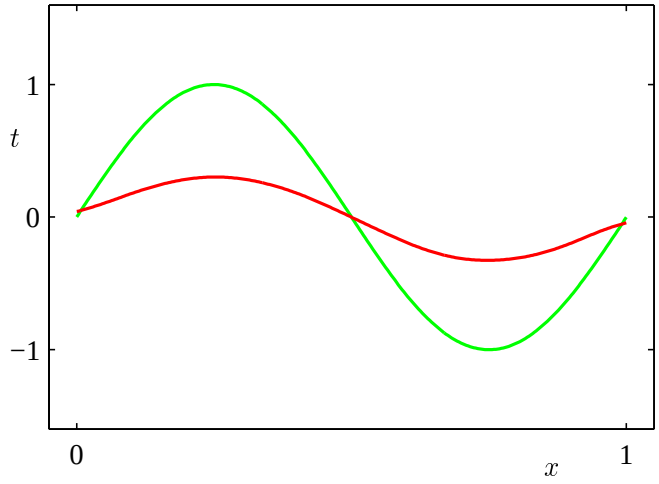
We consider the sinusoidal data set. We generate 100 data sets, each containing $N = 25$ data points, independently from the sinusoidal curve $h(x) = \mathbb{E}\{y|x\} = sin(2\pi x)$. The data sets are indexed by $l = 1, \ldots, L$ where $L = 100$. For each data set $\mathcal{D}^{(l)}$, we fit a model with 24 Gaussian basis functions through a method of ridge regression to give a prediction function $f^{(l)}(x)$.

$$\begin{aligned} \overline{f}(x) &= \frac{1}{L}\sum_{l=1}^{L} f^{(l)}(x), \\ (\text{bias})^2 &= \frac{1}{N}\sum_{t=1}^{N}\left[\overline{f}(x_t) - h(x_t)\right]^2, \\ \text{variance} &= \frac{1}{N}\sum_{t=1}^{N}\frac{1}{L}\sum_{l=1}^{L}\left[f^{(l)}(x_t) - \overline{f}(x_t)\right]^2. \end{aligned}$$

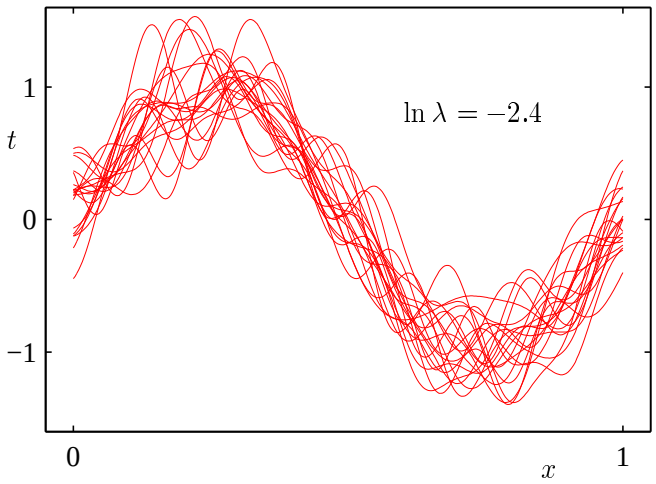
Rigid model (1st row) and flexible model (2nd row)



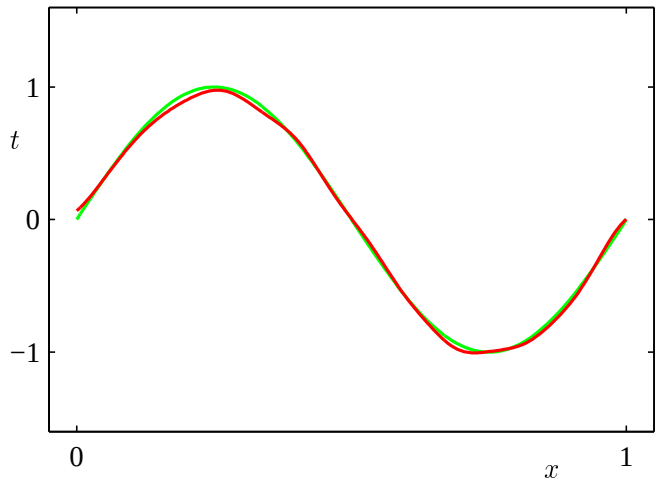
low variance



high bias



high variance



low bias

Bayesian Linear Regression

The likelihood function is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}) = \prod_{t=1}^{N} p(y_t \mid \boldsymbol{\phi}_t, \boldsymbol{w}),$$

where $p(y_t \mid \boldsymbol{\phi}_t, \boldsymbol{w}) = \mathcal{N}(y_t \mid \boldsymbol{w}^\top \boldsymbol{\phi}_t, \beta^{-1})$.
Assuming Gaussian prior for $\boldsymbol{w}$, i.e., $p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$, the posterior distribution over $\boldsymbol{w}$ is again Gaussian that is of the form

$$p(\boldsymbol{w} \mid \boldsymbol{y}, \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$

where

$$\begin{aligned} \boldsymbol{\mu}_N &= \boldsymbol{\Sigma}_N(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\mu}_0 + \beta\boldsymbol{\Phi}^\top \boldsymbol{y}), \\ \boldsymbol{\Sigma}_N^{-1} &= \boldsymbol{\Sigma}_0^{-1} + \beta\boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned}$$

For the sake of simplicity, we consider a particular form of Gaussian prior (zero mean isotropic Gaussian prior),

$$p(\boldsymbol{w} \mid \alpha) = \mathcal{N}(\boldsymbol{w} \mid 0, \alpha^{-1}\boldsymbol{I}).$$

Then the corresponding posterior distribution over $\boldsymbol{w}$ is given by

$$p(\boldsymbol{w} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{w} \mid \boldsymbol{\mu}_N, \boldsymbol{\Sigma}_N),$$

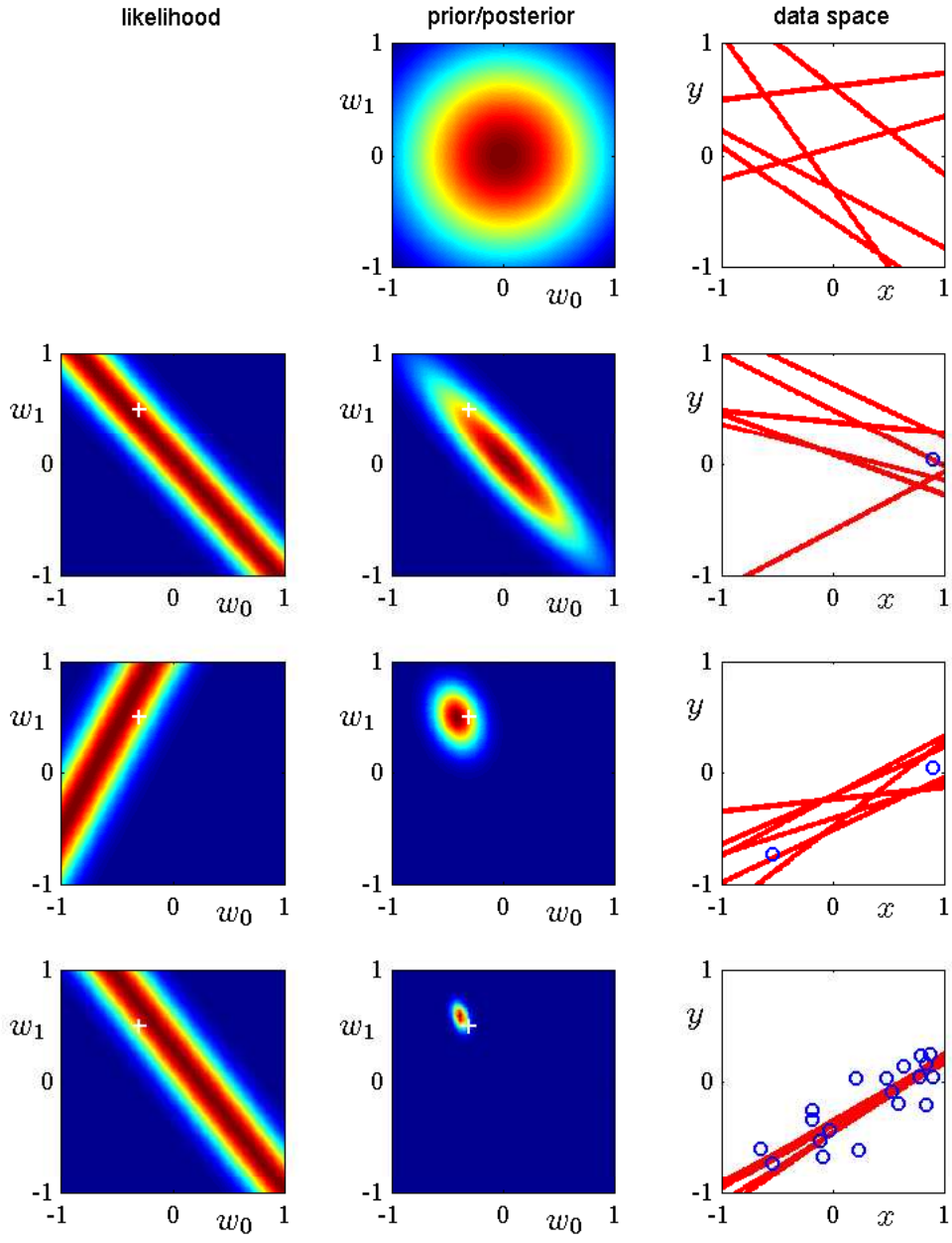
where

$$\begin{aligned} \boldsymbol{\mu}_N &= \beta \boldsymbol{\Sigma}_N \boldsymbol{\Phi}^\top \boldsymbol{y}, \\ \boldsymbol{\Sigma}_N^{-1} &= \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}. \end{aligned}$$

An Example of Sequential Bayesian Learning

Consider a linear model $y = w_0 + w_1 x$ where only two parameters are involved. Next slide illustrates the sequential nature of Bayesian learning, showing that the posterior distribution over parameters become sharper as more data points are observed.

- Left-hand column: likelihood, $p(y|x, \boldsymbol{w})$.
- Middle column: prior/posterior, $p(\boldsymbol{w}|D_t)$.
- Right-hand column: samples of function $y = w_0 + w_1 x$ where $\boldsymbol{w}$ are drawn from the posterior.


likelihood
prior/posterior
data space
w_0
w_1
x
y

Predictive Distribution

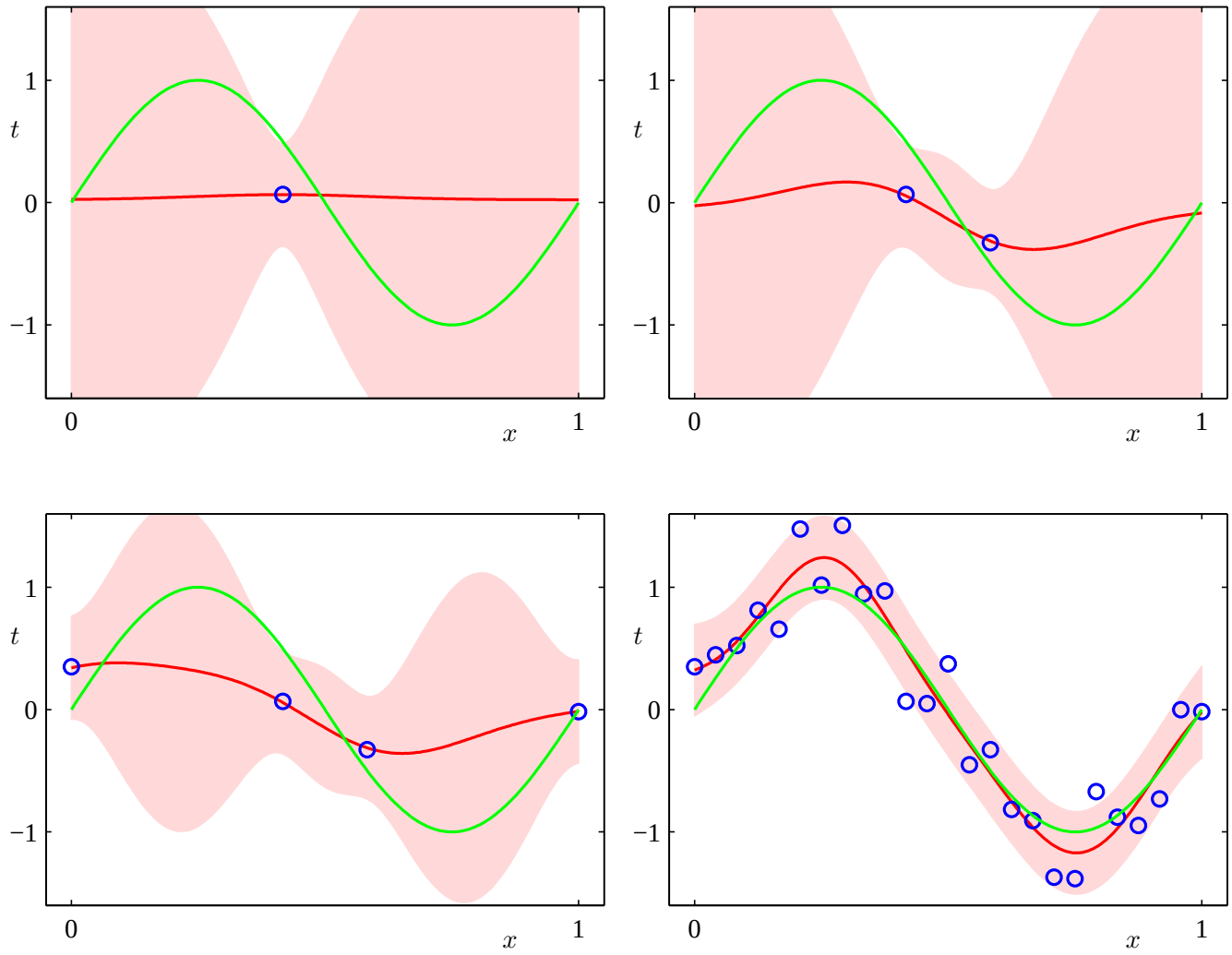
Make predictions of y_* for new values $\boldsymbol{x}_*$ (i.e, $\boldsymbol{\phi}_*$). Let $\mathcal{D} = \{\boldsymbol{\Phi}, \boldsymbol{y}\}$. Then the predictive distribution is given by

$$\begin{aligned} p(y_* \mid \boldsymbol{\phi}_*, \mathcal{D}, \alpha, \beta) &= \int p(y_* \mid \boldsymbol{\phi}_*, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \mathcal{D}, \alpha, \beta) d\boldsymbol{w} \\ &= \mathcal{N}\left(y_* \mid \boldsymbol{\mu}_N^\top \boldsymbol{\phi}_*, \sigma_N^2(\boldsymbol{x}_*)\right), \end{aligned}$$

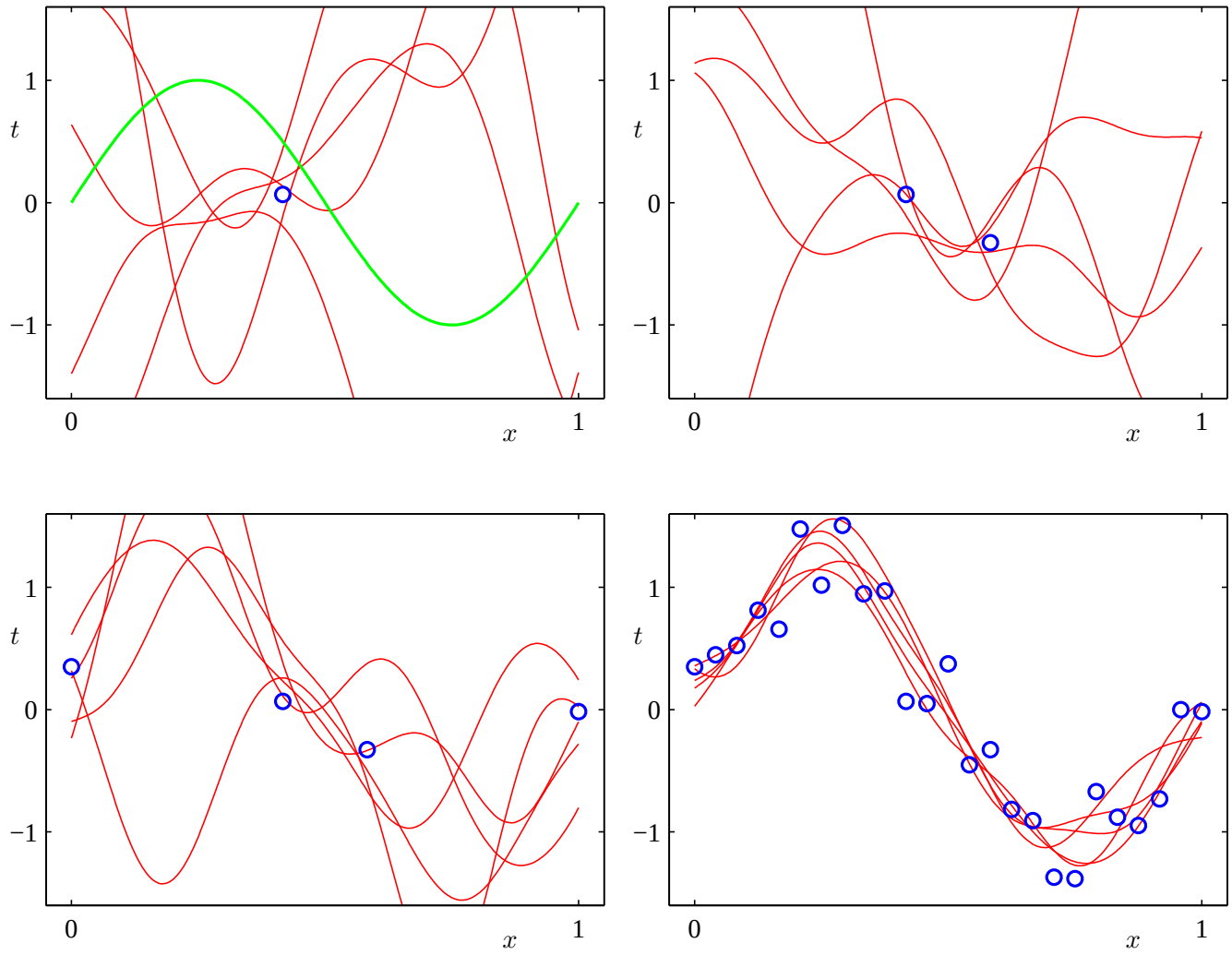
where

$$\begin{aligned} \boldsymbol{\mu}_N &= \beta \boldsymbol{\Sigma}_N \boldsymbol{\Phi}^\top \boldsymbol{y}, \quad \boldsymbol{\Sigma}_N^{-1} = \alpha \boldsymbol{I} + \beta \boldsymbol{\Phi}^\top \boldsymbol{\Phi}, \\ \sigma_N^2(\boldsymbol{x}_*) &= \underbrace{\frac{1}{\beta}}_{\text{noise on the data}} + \underbrace{\boldsymbol{\phi}_*^\top \boldsymbol{\Sigma}_N \boldsymbol{\phi}_*}_{\text{uncertainty associated with parameters } \boldsymbol{w}}. \end{aligned}$$

Examples of Predictive Distributions



Plots of $f(x; \boldsymbol{w})$ using samples from $p(\boldsymbol{w}|\boldsymbol{y})$



Bayesian Model Comparison

- Avoid the over-fitting associated with maximum likelihood by marginalizing over the model parameters instead of making point estimates of their values.
- Models can be compared directly on the training data, without the need for a validation set.
- Avoids the multiple training runs for each model associated with cross-validation.

Suppose that we wish to compare a set of L models, $\{\mathcal{M}_i\}$ for $i = 1, \ldots, L$. (a model refers to a probability distribution over the observed data $\mathcal{D}$)

Given a training set $\mathcal{D}$, we then wish to evaluate the posterior distribution

$$p(\mathcal{M}_i|\mathcal{D}) \propto \underbrace{p(\mathcal{M}_i)}_{\text{prior}} \underbrace{p(\mathcal{D}|\mathcal{M}_i)}_{\text{evidence}}.$$

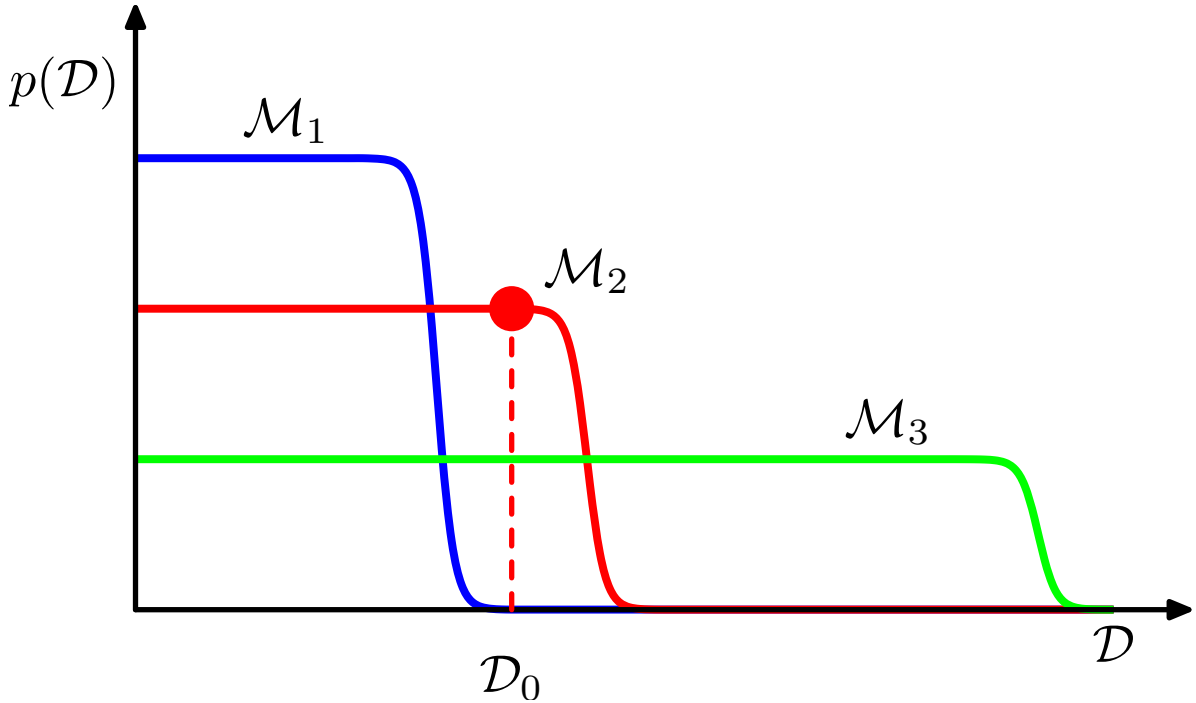
- $p(\mathcal{M}_i)$ is a prior probability distribution to express our uncertainty. The data is generated from one of these models but we are uncertain which one.
- $p(\mathcal{D}|\mathcal{M}_i)$ is the model evidence which expresses the preference shown by the data for different models. This is also known as marginal likelihood, since it can be viewed as a likelihood function over the space of models, in which the parameters have been marginalized out.

- Bayes factor is the ratio of model evidences for two models:

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

- Model selection: Choose the single most probable model. For a model governed by a set of parameters $\boldsymbol{w}$, the model evidence is given by

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



- $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex.
- For the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.

Marginal Likelihood: Hyperparameter Estimation

We estimate hyperparameters α and β through maximizing the marginal likelihood.
The marginal likelihood is given by

$$p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \alpha, \beta) = \int p(\boldsymbol{y} \mid \boldsymbol{\Phi}, \boldsymbol{w}, \beta) p(\boldsymbol{w} \mid \alpha) d\boldsymbol{w}.$$

Marginal likelihood maximization is illustrated in detail in Sec. 3.5.1 and 3.5.2.