Linear Models for Regression

A self-study metarials for PRML [1]

Jisung Lim[1]

[1] B.S. Candidate of Industrial Engineering
Yonsei University, South Korea.

8th February, 2017


Christopher M. Bishop. **Pattern Recognition and Machine Learning**. Springer, 2006.

Summary

Main Concept of Linear Regression

The main concept of linear regression

The goal of regression is to predict the value of one or more continuous target varaible t (or $\boldsymbol{t}$ for more than one variable) given the value of a D-dimensional vector $\boldsymbol{x}$ of input variables. Among them, linear regression model is a broad class of functions which share the property of being linear functions with respect to the adjustable parameters θ_j.

Linear Models and Regression

Three componenets of linear regression

We may see the linear regression as a process comprise of three components

1. **N observations**
   At first, We retrieve a training data set of N observations $\boldsymbol{x}_n$ together with corresponding target values t_n.
   $$\mathcal{D} = \{(t_1, \boldsymbol{x}_1), \ldots, (t_N, \boldsymbol{x}_N)\} \tag{1}$$
2. **Linear Models for Regression**
   Given N data set, we will construct linear regression models that comprise a broad class of functions which share the property of being linear functions with respect to the adjustable parameters θ_j. We may extend the class of models by using basis functions $\phi_j(\cdot)$.
   $$y(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{2}$$
3. **Prediction**
   Then, we can construct a function $y(\boldsymbol{x})$ which constitute the predictions for the target value t.
   $$\hat{t} = y(\boldsymbol{x}) \tag{3}$$
   or more generally, from a probabilistic perspective we can construct probability distribution of t given $\boldsymbol{x}$
   $$p(t|\boldsymbol{x}) \tag{4}$$
   which expresses our uncertainty about the value of t for each value of $\boldsymbol{x}$.

The Form of Linear Models

Linear regression model is a broad class of functions which share the property of being linear functions of the adjustable parameters θ_j. The form of linear regression models varies from the form of linear combination of parameters θ_j and input variables x_j to the form of linear combination of the parameters and a fixed set of nonlinear functions.

Given $\boldsymbol{x} = (x_1, \ldots, x_D)^T$, the simplest linear model for regression forms:

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{D} \theta_j x_j \tag{5}$$

which is often simply known as *linear regression*. When augmented with other forms of basis function expansion, it can model also nonlinear relationships.

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \theta_0 + \sum_{j=1}^{M-1} \theta_k \phi_j(\boldsymbol{x}) \tag{6}$$

or equivalently, introducing dummy basis function $\phi_0(\boldsymbol{x}) = 1$ so that

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \boldsymbol{\theta}^T \boldsymbol{\phi}(\boldsymbol{x}) \tag{7}$$

the fixed set of nonlinear functions of the input variables $\boldsymbol{\phi}(\boldsymbol{x})$ is known as *basis functions* and θ_0 as *bias parameter* which allows for any fixed offset in the data.

Linear Models and Basis Function

In machine learning approach, we will apply some form of fixed pre-processing or feature extraction, to the original data variables.

$$\begin{aligned} \text{original variables}: \ & \boldsymbol{x} = (x_1, \ldots, x_N)^T \\ \text{features}: \ & \boldsymbol{\phi}(\boldsymbol{x}) = (\phi_0(\boldsymbol{x}), \ldots, \phi_{M-1}(\boldsymbol{x}))^T \end{aligned} \tag{8}$$

Choose of basis function

- Polynomial basis function $(\text{ex})\ \phi_j(\boldsymbol{x}) = x^j$
  Limitation of polynomial basis function is that they are global functions of the input variables.
- Spline function $(\text{ex})\ \phi_j(\boldsymbol{x}) = f_i(\boldsymbol{x})$ for $\boldsymbol{x} \in \mathcal{R}_i$
  This can be resolved by dividing the input space up into regions and fit a different polynomial $f_i(\boldsymbol{x})$ in each region $\mathcal{R}_i$.
- Gaussian basis function $(\text{ex})\ \phi_j(x) = \exp\{-(x - \mu^2)^2/2s^2\}$
  A probabilistic interpretations are not required, and in particular the normalization coefficient is unimportant because of the parameters w_j which is to be mulitplied to each basis function.
- Sigmoid and $\tanh$ basis function $(\text{ex})\ \phi_j(x) = \sigma((x - \mu_j)/s)$ where $\sigma(a)$ is the logistic sigmoid function. Since $\tanh(a) = 2\sigma(a) - 1$, generalized linear combination of two basis functions can be identically expressed.
- Fourier basis function (ex) wavelets.

Gaussian Noise Assumption

setting 1. Additive gaussian noise
Let us assume that the target variable t is given by a deterministic function $y(\boldsymbol{x}, \boldsymbol{w})$ with additive Gaussian noise so that

$$t = y(\boldsymbol{x}, \boldsymbol{w}) + \epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(\epsilon \mid 0, \beta^{-1}) \tag{9}$$

Then, we may express the distribution of target value t as follows

$$p(t \mid \boldsymbol{x}, \boldsymbol{w}, \beta) = \mathcal{N}(\epsilon \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \tag{10}$$

setting 2. Squared loss function
If we set the loss function as a squared loss function as follows

$$L(t, y(\boldsymbol{x}, \boldsymbol{w})) = \{t - y(\boldsymbol{x}, \boldsymbol{w})\}^2 \tag{11}$$

and also we can define the expected loss as follows

$$\mathbb{E}[L] = \iint L(t, y(\boldsymbol{x}, \boldsymbol{w}))p(t, \boldsymbol{x})\,\mathrm{d}t\,\mathrm{d}\mathbf{x} = \iint \{t - y(\boldsymbol{x}, \boldsymbol{w})\}^2 p(t, \boldsymbol{x})\,\mathrm{d}t\,\mathrm{d}\mathbf{x} \tag{12}$$

Now, to find optimal prediction y, set variational derivative of $\mathbb{E}[L]$ with respect to $y(\boldsymbol{x}, \boldsymbol{w})$ equals to zero $\delta\mathbb{E}[L]/\delta y = 0$, then we get $y^* = \mathbb{E}[t|\boldsymbol{x}]$ which will be simply

$$y^* = \mathbb{E}[t|\boldsymbol{x}] = \int t p(t|\boldsymbol{x})\,\mathrm{d}t = y(\boldsymbol{x}, \boldsymbol{w}) \tag{13}$$

Maximum Likelihood with Basis Function

Now, we will construct predictive function $y(t|\boldsymbol{x}, \boldsymbol{w}, \beta)$ by using maximum likelihood estimation.

N Observations

At first, Lets consider the input data of N observations,

$$\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T \quad \text{and} \quad \mathbf{t} = (t_1, \ldots, t_N)^T \tag{14}$$

Find likelihood function

With i.i.d. assumption, we can get the following likelihood function,

$$L(\boldsymbol{w}) = p(\mathbf{t} \mid \mathbf{X}, \boldsymbol{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \tag{15}$$

taking logarithm of both sides,

$$\begin{aligned} \ln L(\boldsymbol{w}) &= \sum_{n=1}^{N} \ln[\mathcal{N}(t_n \mid \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1})] \\ &= \frac{N}{2} \ln(\beta) - \frac{N}{2} \ln(2\pi) - \frac{\beta}{2} \sum_{n=1}^{N} \{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \\ &= \frac{N}{2} \ln(\beta) - \frac{N}{2} \ln(2\pi) - \beta E_D(\boldsymbol{w}) \end{aligned} \tag{16}$$

where $E_D(\boldsymbol{w})$ is the sum-of-squares error function.

Maximum likelihood with respect to w

Maximize $L(w)$ with respect to w
As we've seen before, maximizing the likelihood function $L(\boldsymbol{w})$ is equivalent to minimizing a sum-of-squares error function $E_D(\boldsymbol{w})$ from (16), given the following assumptions

$$\mathrm{t} \sim \mathcal{N}(t \mid y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1}) \quad \text{with} \quad \text{i.i.d. condition} \tag{17}$$

where $y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)$. Or we may find same result with the gradient of the log likelihood function,

$$\nabla p(\mathbf{t} \mid \boldsymbol{w}, \beta) = \sum_{n=1}^{N} [\{t_n - \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n)^T] \tag{18}$$

Setting the gradient to zero gives

$$0 = \sum_{n=1}^{N} t_n \boldsymbol{\phi}(\boldsymbol{x}_n)^T - \boldsymbol{w}^T \left(\sum_{n=1}^{N} \boldsymbol{\phi}(\boldsymbol{x}_n) \boldsymbol{\phi}(\boldsymbol{x}_n)^T \right) \tag{19}$$

Solving for $\boldsymbol{w}$, we obtain

$$\boldsymbol{w}_{\mathrm{ML}} = (\boldsymbol{\Phi}^T \boldsymbol{\Phi})^{-1} \boldsymbol{\Phi}^T \mathbf{t} \tag{20}$$

Maximum likelihood with respect to β

Maximize $L(\boldsymbol{w})$ with respect to β
We can also maximize the log likelihood function with respect to the noise precision parameter β, setting derivate of likelihood function with respect to β equal to zero as follows

$$\frac{\partial}{\partial \beta}[p(\mathbf{t} \mid \boldsymbol{w}, \beta)] = \frac{N}{2\beta} - E_D(\boldsymbol{w}) = 0 \tag{21}$$

Then we get,

$$\frac{1}{\beta_{\mathrm{ML}}} = \frac{2}{N} E_D(\boldsymbol{w}) = \frac{1}{N} \sum_{n=1}^{N} \left\{ t_n - \boldsymbol{w}_{\mathrm{ML}}^T \boldsymbol{\phi}(\boldsymbol{x}_n) \right\}^2 \tag{22}$$

Predictive Function
Hence, we get predictive function which is given by

$$p(t|\boldsymbol{x}, \boldsymbol{w}_{\mathrm{ML}}, \beta_{\mathrm{ML}}) = \left(\frac{\beta_{\mathrm{ML}}}{2\pi} \right)^{1/2} \exp\left[-\frac{\beta_{\mathrm{ML}}}{2} \left\{ t - \boldsymbol{w}_{\mathrm{ML}}^T \boldsymbol{\phi}(\boldsymbol{x}) \right\}^2 \right] \tag{23}$$

Normal Equation for the least squares problem

Normal Equation

Now consider some conventional terms with the solution (20) for $\boldsymbol{w}$

$$\boldsymbol{w}_{\mathrm{ML}} = \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T\mathbf{t}$$

which are known as the *normal equations* for the least squares problem.

Design Matrix

Here $\boldsymbol{\Phi}$ is an $N \times M$ matrix, called the *design matrix*,

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^T \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_N)^T \end{bmatrix} = \begin{bmatrix} \boldsymbol{\varphi}_0 & \cdots & \boldsymbol{\varphi}_{M-1} \end{bmatrix} \tag{24}$$

where

$$\begin{aligned} \boldsymbol{\phi}(\boldsymbol{x}_n) &= \left(\phi_0(\boldsymbol{x}_n), \ldots, \phi_{M-1}(\boldsymbol{x}_n)\right)^T \\ \boldsymbol{\varphi}_j &= \left(\phi_j(\boldsymbol{x}_1), \ldots, \phi_j(\boldsymbol{x}_N)\right)^T \end{aligned} \tag{25}$$

Moore-Penrose pseudo-inverse

And then, the quantity

$$\boldsymbol{\Phi}^{\dagger} \equiv \left(\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^T \tag{26}$$

is known as the *Moore-Penrose pseudo-inverse* of the matrix $\boldsymbol{\Phi}$.

The Role of Bias Parameter w_0

We can get some insight of the role of *Bias Parameter* w_0 by solving the equation

$$\frac{\mathrm{d}}{\mathrm{d}w_0} E_D(\boldsymbol{w}) = 0 \tag{27}$$

At first, let us make the bias parameter to be explicit.

$$E_D(\boldsymbol{w}) = \frac{1}{2} \sum_{n=1}^{N} \left\{ t_n - w_0 - \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x}_n) \right\}^2 \tag{28}$$

Setting the derivative with respect to w_0 equal to zero, and solving for w_0, we obtain

$$\begin{aligned} w_0 &= \frac{1}{N} \sum_{n=1}^{N} t_n - \sum_{j=1}^{M-1} [w_j \frac{1}{N} \sum_{n=1}^{N} \phi_j(\boldsymbol{x}_n)] \\ &= \bar{t} - \sum_{j=1}^{M-1} w_j \bar{\phi}_j \end{aligned} \tag{29}$$

Thus, the bias w_0, as a kind of offset, compensates for the gap between the averages of the target values $\bar{t}$ and the linear combination of adaptive paramters w_j and the average of the basis functions $\bar{\phi}_j$.

Geometry of least squares

Geometry of least squares

Let us consider an N-dimensional space whose axes are given by the t_n, so that $\mathbf{t} = (t_1, \ldots, t_N)^T$ is a specific vector in this space.

$$\mathbf{t} = (t_1, \ldots, t_N)^T \in \mathbb{R}^N \tag{30}$$

Each basis function $\phi_j(\boldsymbol{x}_n)$ may be represented as a vector $\boldsymbol{\varphi}_j$ in the same space, when evaluated at the N data points.

$$\boldsymbol{\varphi}_j \in \mathbb{R}^N \quad \text{where} \quad \boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}_0 & \cdots & \boldsymbol{\varphi}_{M-1} \end{bmatrix} \tag{31}$$

Since $\mathbf{y}$ is an arbitrary linear combination of the vector $\boldsymbol{\varphi}_j$, it lies on the M-dimensional subspace $\mathcal{S}$.

$$\mathbf{y} = \boldsymbol{\Phi}\boldsymbol{w} = \begin{bmatrix} \boldsymbol{\varphi}_0 & \cdots & \boldsymbol{\varphi}_{M-1} \end{bmatrix} \begin{bmatrix} w_0 \\ \vdots \\ w_{M-1} \end{bmatrix} = \sum_{j=0}^{M-1} w_j \boldsymbol{\varphi}_j \tag{32}$$

Hence, intutively, the sum-of-squares error may be understood as the squared Euclidean distance between $\mathbf{y}$ and $\mathbf{t}$. Thus the least-squares solution for $\boldsymbol{w}$ corresponds to the choice of $\mathbf{y}$ that lies in subspace $\mathcal{S}$ and that is closest to $\mathbf{t}$.

Sequential learning

In cases where the training data set is very large or data is received in a stream, a direct solution using the normal equations may not be possible. An alternative approach is to use sequential algorithms, also known as on-line algorithms. Sequential learning is appropriate for realtime applications in which the data observations are arriving in a continuous steam, and predictions must be made before all of the data points are seen. At first, we will consider tha data points $D_n = (n = 1, 2, 3, \ldots)$ one at a time and then update model parameters after each such consideration.

$$\mathcal{D} = \{D_1, \ldots, D_n, \ldots\} = \{(t_1, \boldsymbol{x}_1), \ldots, (t_n, \boldsymbol{x}_n), \ldots\} \tag{33}$$

After presentation of D_n, the stochastic gradient descent algorithm updates the parameter vector $\boldsymbol{w}$ using

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \nabla E_n \tag{34}$$

The total error function E is the sum of a given error function E_n, evaluated from the n-th data point D_n.

$$E = \sum_{\forall n} E_n \quad \text{where} \quad E_n = d(t_n, y(\boldsymbol{x}_n)) = \{t_n - \boldsymbol{w}^{(\tau)T} \boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{35}$$

and the algorithms is given as follows, where η rerepsents a learning rate.

$$\boldsymbol{w}^{(\tau+1)} = \boldsymbol{w}^{(\tau)} - \eta \{t_n - \boldsymbol{w}^{(\tau)T} \boldsymbol{\phi}(\boldsymbol{x}_n)\} \boldsymbol{\phi}(\boldsymbol{x}_n) \tag{36}$$

This is known as *least-mean-squares* or the *LMS algorithm*.

Regularized Least Squares

Let us adding a regularization term to an error function in order to prevent model from over-fitted. The total error function to be minimized takes the form

$$E_D(\boldsymbol{w}) + \lambda E_W(\boldsymbol{w}) \tag{37}$$

where λ is the regulariation coefficient that controls the relative importance between the data-dependent error $E_D(\boldsymbol{w})$ and the regularization term $E_W(\boldsymbol{w})$.

One of the simple forms of regularizer is given by the sum-of-squares of the weight vector elements

$$E_W(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{38}$$

If we consider the sum-of-squares error function given by

$$E_D(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right\}^2 \tag{39}$$

then the total error function becomes

$$E(\boldsymbol{w}) = E_D(\boldsymbol{w}) + E_W(\boldsymbol{w}) = \frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\right\}^2 + \frac{\lambda}{2}\boldsymbol{w}^T\boldsymbol{w}. \tag{40}$$

Other choices of regularizer

Generalized regularizer

A more general regularizer is sometimes used, for which the regularized error takes the form

$$\frac{1}{2}\sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\right\}^2 + \frac{\lambda}{2}\sum_{j=1}^{M}|w_j|^q \tag{41}$$

where $q = 2$ corresponds to the quadratic regularizer. There are conventional terms for linear and quadratic regularizers, respectively, *lasso* and *ridge*.

Minimizing regularized model and constraind optimization

For sufficiently large $\lambda = \lambda^*$, we have to minimize regularized model (38) to obtain the solution $\boldsymbol{w}^*$.

$$\underset{w_j}{\text{Minimize}} \quad \sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\right\}^2 + \lambda^*\sum_{j=1}^{M}|w_j|^q \tag{42}$$

This optimization problem can be seen as a Lagrange multipliers, so that it can be represented by constraind optimization problem.

$$\begin{aligned} &\underset{w_j}{\text{Minimize}} && \sum_{n=1}^{N}\left\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})\right\}^2 \\ &\text{subject to} && \sum_{j=1}^{M}|w_j|^q \leq \eta^* \end{aligned} \tag{43}$$

Multiple Outputs

A Frequentist Viewpoint of The Model Complexity

Model Complexity

The use of MLE can lead to severe over-fitting if complex models are trained using data sets of limited size. For releasing the over-fitting problem, we've discussed two approaches:

- limiting the number of basis functions
- introducing regularization terms

But, *the first approach* has side effect of limiting the flexibility of the model to capture interesting and important trends in the data. Although, *the second approach* reveals the problem of choosing the appropriate coefficient λ.

A frequentist viewpoint of the model complexity

Now we consider the over-fitting issue and model complexity. As we've seen before, an over-fitting issue doesn't incur when we marginalize over parameters in a Bayesian setting. In the frequentist setting, however, they handle the model complexity issue with a special viewpoint, known as the *bias-variance* trade-off. At first, we will consider the frequentist view point.

Loss Function for Regression

In decision theory, we consider the family of functions $y(\boldsymbol{x})$ which is estimate of the value of t, for each input $\boldsymbol{x}$. Generally, the optimal function $y^*(\boldsymbol{x})$ can be chosen by minimizing the loss function. And a popular choice of the loss function is the squared loss function which is given by

$$L(t, y(\boldsymbol{x})) = \{t - y(\boldsymbol{x})\}^2 \tag{44}$$

and the expected loss function is given by

$$\begin{aligned} \mathbb{E}[L] &= \iint L(t, y(\boldsymbol{x}))p(t, \boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \\ &= \iint \{t - y(\boldsymbol{x})\}^2 p(t, \boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}t \end{aligned} \tag{45}$$

Now we want to find the optimal function $y^*(\boldsymbol{x})$, which minimizes the expected loss (42), and we can get by setting the variational derivative $\delta\mathbb{E}[L]/\delta y(\boldsymbol{x})$ equals to zero as follows

$$\begin{aligned} &\frac{\delta\mathbb{E}[L]}{\delta y(\boldsymbol{x})} = 2\int \{y(\boldsymbol{x}) - t\}p(\boldsymbol{x}, t)\,\mathrm{d}t = 0 \\ &\iff y(\boldsymbol{x})\int p(\boldsymbol{x}, t)\mathrm{d}t = \int tp(\boldsymbol{x}, t)\mathrm{d}t \\ &\iff y(\boldsymbol{x}) = \int t\frac{p(\boldsymbol{x}, t)}{p(\boldsymbol{x})}\mathrm{d}t = \int tp(t|\boldsymbol{x})\mathrm{d}t = \mathbb{E}[t|\boldsymbol{x}] \end{aligned} \tag{46}$$

Loss Function for Regression (conti. . .)

From the result of (43), for convenience, denote the conditional expectation by $h(\boldsymbol{x})$, which is given by

$$h(\boldsymbol{x}) = \mathbb{E}[t|\boldsymbol{x}] = \int tp(t|\boldsymbol{x})\mathrm{d}t \tag{47}$$

Now the expected loss function can be written in the form

$$\begin{aligned}\mathbb{E}[L] &= \mathbb{E}[\{t - h(\boldsymbol{x})\}^2] + \mathbb{E}[\{h(\boldsymbol{x}) - y(\boldsymbol{x})\}^2] + \mathbb{E}[2\{t - h(\boldsymbol{x})\}\{h(\boldsymbol{x}) - y(\boldsymbol{x})\}] \\ &= \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\,\mathrm{d}\mathbf{x} + \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t)\,\mathrm{d}\mathbf{x}\mathrm{d}t\end{aligned} \tag{48}$$

since

$$\begin{aligned}\mathbb{E}[\{t - h(\boldsymbol{x})\}\{h(\boldsymbol{x}) - y(\boldsymbol{x})\}] &= \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\} \int \{h(\boldsymbol{x}) - t\} p(\boldsymbol{x}, t)\,\mathrm{d}t\,\mathrm{d}\boldsymbol{x} \\ &= \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\} \left\{ h(\boldsymbol{x})p(\boldsymbol{x}) - \int tp(\boldsymbol{x}, t)\mathrm{d}t \right\} \mathrm{d}\boldsymbol{x} \\ &= 0 \quad \left(\because h(\boldsymbol{x}) = \int tp(t|\boldsymbol{x})\mathrm{d}t \right)\end{aligned} \tag{49}$$

Loss Function for Regression (conti. . .)

Now consider the equation (45)

$$\mathbb{E}[L] = \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\,\mathrm{d}\mathbf{x} + \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t)\,\mathrm{d}\mathbf{x}\mathrm{d}t$$

The second term: intrinsic, invariant noise

$$\mathbb{E}[\{h(\boldsymbol{x}) - t\}^2] = \iint \{h(\boldsymbol{x}) - t\}^2 p(\boldsymbol{x}, t)\,\mathrm{d}\mathbf{x}\mathrm{d}t$$

- The second term is independent of $y(\boldsymbol{x})$.
- This quantity arises from the intrinsic noise on the data.
- It represents the minimum achievable value of the expected loss.

The first term: Error arose from the choice for the function $y(\boldsymbol{x})$

$$\mathbb{E}[\{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2] = \int \{y(\boldsymbol{x}) - h(\boldsymbol{x})\}^2 p(\boldsymbol{x})\,\mathrm{d}\mathbf{x}$$

- The first term depends on our choice for the function $y(\boldsymbol{x})$.
- We will seek a solution for $y(\boldsymbol{x})$ which makes this term a minimum.
- Because it is nonnegative, the smallest that we can hope to make this term is zero.
- If the data are sufficient enough, we could find the regression function $h(\boldsymbol{x})$ in any desired degree of accuracy.

Frequentist treatment for modeling the $h(\boldsymbol{x})$

In a veiwpoint of frequentist, when we model the function $h(\boldsymbol{x})$ using the parametric function $y(\boldsymbol{x}, \boldsymbol{w})$, since the frequentist treatment involves making a point estiamte of $\boldsymbol{w}$ based on the data set $\mathcal{D}$, they try to understand the uncertainty of the estimate as the frequency when we suppose a large number of data sets each of which satisfies i.i.d. condition.

Let us consider the case where the size of data set is N, the number of data set is L, and each data set is drawn from the distribution $p(t, \boldsymbol{x})$ independently.

$$\mathcal{D}^{(l)} = \{(t_1, \boldsymbol{x}_1), \ldots, (t_N, \boldsymbol{x}_N)\} \quad \text{where} \quad l = 1, \ldots, L \quad \text{and} \quad \mathcal{D}^{(l)} \sim p(t, \boldsymbol{x}) \tag{50}$$

Then we could run the learning algorithm to obtain a prediction function $y(\boldsymbol{x}; \mathcal{D})$ and then, from the first term of the equation (45), the integrand will be given by

$$\{y(\boldsymbol{x}; \mathcal{D}) - h(\boldsymbol{x})\}^2 = \{(y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}[y(\boldsymbol{x}; \mathcal{D})]) + (\mathbb{E}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x}))\}^2 \tag{51}$$

for any given specific data set $\mathcal{D}$.

Frequentist treatment for modeling the $h(x)$ (conti. . .)

Since the quantity (48) will be dependent on the particular data set $\mathcal{D}$, we should take its average over all data sets.

$$\begin{aligned}\mathbb{E}[\{y(\boldsymbol{x};\mathcal{D})-h(\boldsymbol{x})\}^2] &= \mathbb{E}\left[\{(y(\boldsymbol{x};\mathcal{D})-\mathbb{E}[y(\boldsymbol{x};\mathcal{D})])+(\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x}))\}^2\right] \\ &= \{\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x})\}^2+\mathbb{E}\left[\{y(\boldsymbol{x};\mathcal{D})-\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]\}^2\right]\end{aligned} \tag{52}$$

since

$$\mathbb{E}[\{y(\boldsymbol{x};\mathcal{D})-\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]\}\{\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x})\}]=0 \tag{53}$$

Hence, the average quantity over all data sets is given by the form

$$\mathbb{E}[\{y(\boldsymbol{x};\mathcal{D})-h(\boldsymbol{x})\}^2] = \underbrace{\{\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]-h(\boldsymbol{x})\}^2}_{(\text{bias})^2}+\underbrace{\mathbb{E}\left[\{y(\boldsymbol{x};\mathcal{D})-\mathbb{E}[y(\boldsymbol{x};\mathcal{D})]\}^2\right]}_{\text{variance}} \tag{54}$$

where the equation might be interpreted as the **bias-variance tradeoffs**.

Frequentist treatment for modeling the $h(x)$ (conti. . .)

Hence, now we can consider the expected loss under the frequentist setting as follows

$$\text{expected loss} = (\text{bias})^2 + \text{variance} + \text{noise} \tag{55}$$

$$\textit{where} \quad (\text{bias})^2 = \int \{\mathbb{E}[y(\boldsymbol{x}; \mathcal{D})] - h(\boldsymbol{x})\}^2 p(\boldsymbol{x}) \mathrm{d}\mathbf{x}$$

$$\text{variance} = \int \{y(\boldsymbol{x}; \mathcal{D}) - \mathbb{E}[y(\boldsymbol{x}; \mathcal{D})]\}^2 p(\boldsymbol{x}) \mathrm{d}\mathbf{x} \tag{56}$$

$$\text{noise} = \iint \{h(\boldsymbol{x}) - t\}^2 p(t, \boldsymbol{x}) \mathrm{d}\boldsymbol{x} \mathrm{d}t$$

Since we only have finite data sets where $l = 1, \ldots, L$, the bias and variance terms are given by

$$(\text{bias})^2 = \frac{1}{N} \sum_{n=1}^{N} \{\bar{y}(\boldsymbol{x}_n) - h(\boldsymbol{x}_n)\}^2$$

$$\text{variance} = \frac{1}{N} \sum_{n=1}^{N} \frac{1}{L} \sum_{l=1}^{L} \{y^{(l)}(\boldsymbol{x}_n) - \bar{y}(\boldsymbol{x}_n)\}^2 \tag{57}$$

$$\textit{where} \quad \bar{y}(\boldsymbol{x}) = \mathbb{E}[y(\boldsymbol{x}; \mathcal{D})] = \frac{1}{L} \sum_{l=1}^{L} y^{(l)}(\boldsymbol{x})$$

Frequentist treatment for modeling the $h(x)$ (conti. . .)

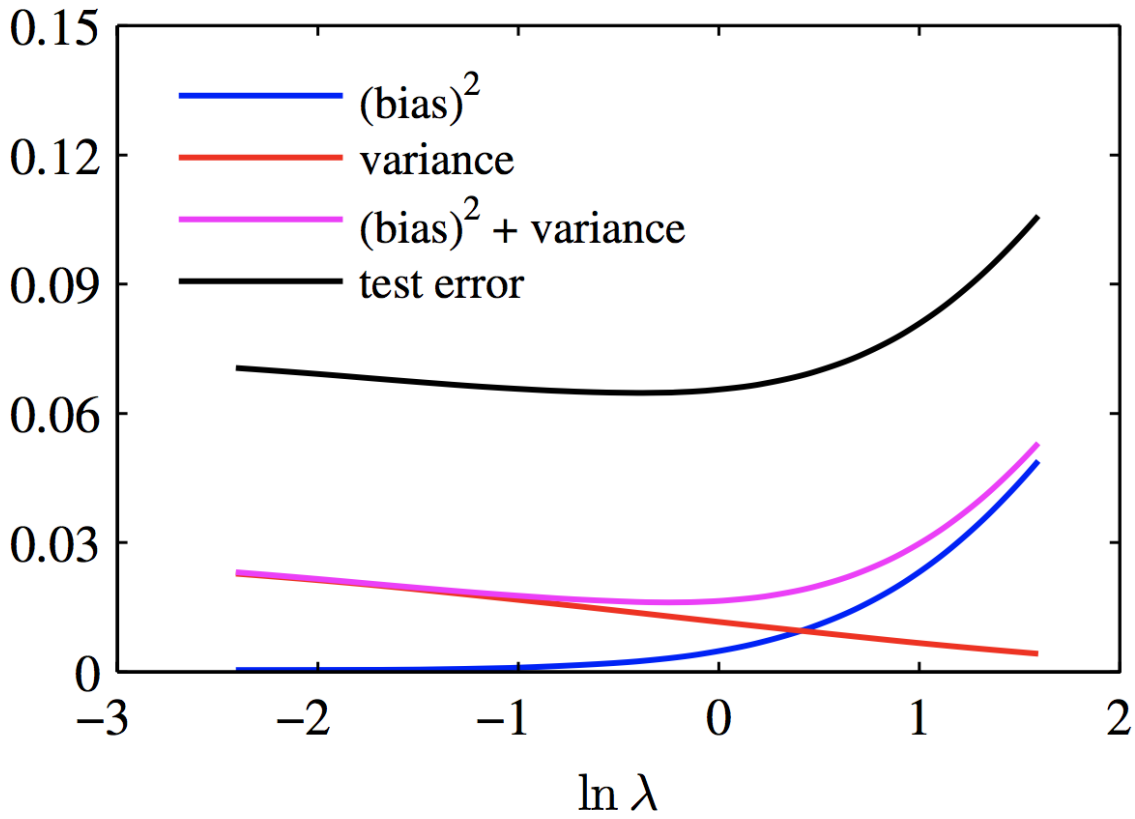

Figure 1: Loss function for the regularized least-square model.

Bayesian Approach in Linear Regression

What we've discussed so far

- Effective model complexity; governed by the number of basis function.
- Regularizer; governed by the value of the regularization coefficient.
- Validation set; too much expensive and wasteful.

We, therefore, turn to a Bayesian treatment of linear regression, which

- will avoid the over-fitting problem of Maximum Likelihood approach.
- will determine the model complexity automatically, using the training data alone.

What we will do

- Formulate our knoweledge about the model itself, and prior belief for the uncertainty of parameters.

$$\text{Model} : y(\boldsymbol{x}, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}) \quad \text{Parameter} : p(\boldsymbol{w}|\alpha) \tag{58}$$

- Observe data of N observations

$$\mathcal{D} = \{(t_1, \boldsymbol{x}_1), \ldots, (t_N, \boldsymbol{x}_N)\} \tag{59}$$

- Evaluate posterior distribution for the parameters

$$p(\boldsymbol{w}|\mathcal{D}) \propto L(\boldsymbol{w}) p(\boldsymbol{w}) \tag{60}$$

- Make prediction

$$p(t|\mathbf{t}, \alpha, \beta) = \int p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta) \mathrm{d}\boldsymbol{w} \tag{61}$$

Bayesian Treatment

First, assume we get the likelihood function from the data which has the form

$$\text{likelihood}: \ p(\mathbf{t}|\boldsymbol{w}) = \prod_{n=1}^{N} \mathcal{N}(t_n|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \sim \text{Gaussian distribution} \tag{62}$$

Since the likelihood function is given by a Gaussian distribution so that the conjugate prior and corresponding posterior is also given by a Gaussain distribution.

$$\text{prior}: \ p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\mathbf{m}_0, \mathbf{S}_0) \tag{63}$$

$$\text{posterior}: \ p(\boldsymbol{w}|\mathbf{t}) = \mathcal{N}(\boldsymbol{w}|\mathbf{m}_N, \mathbf{S}_N) \tag{64}$$

where

$$\begin{aligned} \mathbf{m}_N &= \mathbf{S}_N(\mathbf{S}_0^{-1}\mathbf{m}_0 + \beta\boldsymbol{\Phi}^T\mathbf{t}) \\ \mathbf{S}_N^{-1} &= \mathbf{S}_0^{-1} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \end{aligned} \tag{65}$$

Note that,

- $\boldsymbol{w}_{\text{MAP}} = \mathbf{m}_N$, since the Guassian is unimodal distribution whose mode is coincides with its mean.
- Also if we chose an infinitely broad prior $\mathbf{S}_0 = \alpha^{-1}\mathbf{I}$ with $\alpha \to 0$, the posterior mean $\mathbf{m}_N$ reduces to the value $\boldsymbol{w}_{\text{ML}}$.
- Similarly, if $N = 0$, the posterior and prior are same.
- Furthermore, from the concept of on-line learning, the posterior at any stage would become the prior at the next stage.

Bayesian Treatment

Set the prior
For simplicity, we will set the prior distribution as follows

$$p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \tag{66}$$

Evaluate the posterior
We, therefore, get the mean and variance of the posterior given by

$$\begin{aligned} \mathbf{m}_N &= \beta \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t} \\ \mathbf{S}_N^{-1} &= \alpha \mathbf{I} + \beta \boldsymbol{\Phi}^T \boldsymbol{\Phi} \end{aligned} \tag{67}$$

Maximum a posteriori approach
Since the posterior is proportional to the product of the prior and the likelihood function, the log of the posterior is given by

$$\ln p(\boldsymbol{w}|\mathbf{t}) = -\frac{\beta}{2}\sum_{n=1}^{N}\{t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 - \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} + \text{const.} \tag{68}$$

where maximizing the posterior distribution with respect to $\boldsymbol{w}$ is equivalent to the minimizing the regularized sum-of-squares error function.

Bayesian Treatment

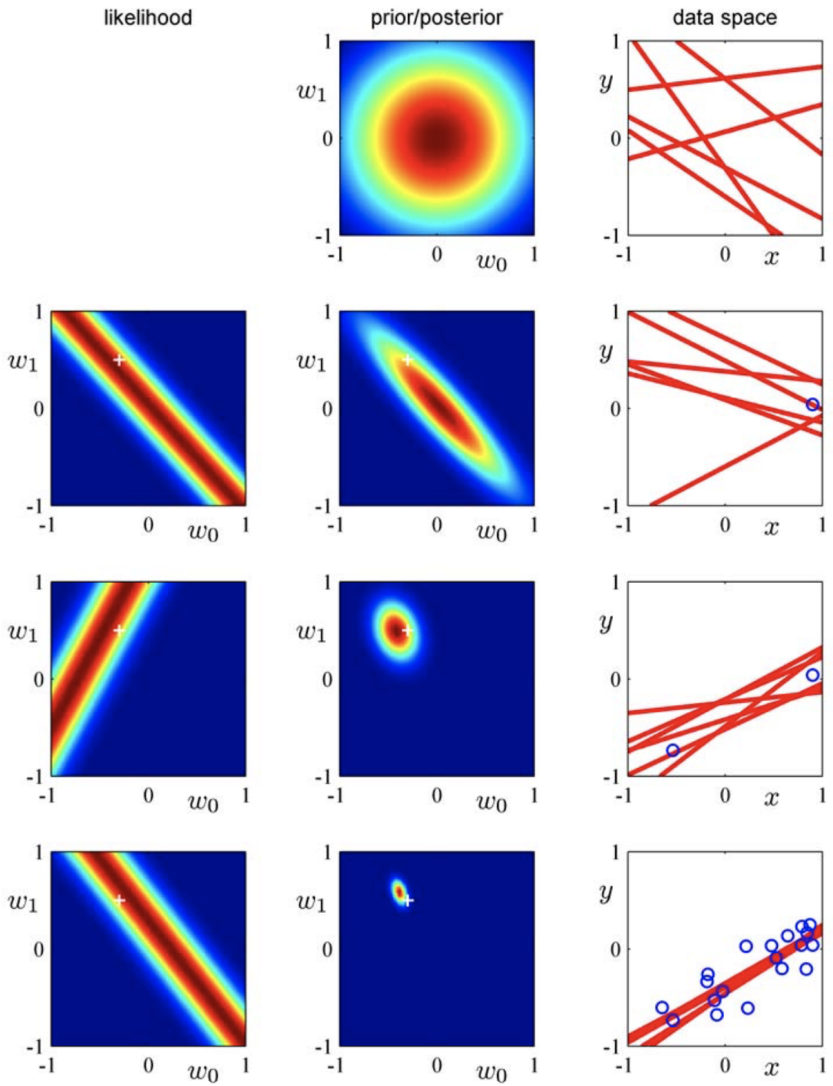

Figure 2: Illustration of the sequential bayesian treatment.

Predictive distribution

Indeed, marginalizing multiple solutions with respect to the distribution of the parameter $\boldsymbol{w}$ lies at the heart of a Bayesian approach. We, therefore, get the predictive distribution given by

$$\begin{aligned} p(t|\mathbf{t}, \alpha, \beta) &= \int p(t, \boldsymbol{w}|\mathbf{t}, \alpha, \beta)\ \mathrm{d}\mathbf{w} \\ &= \int p(t|\boldsymbol{w}, \mathbf{t}, \alpha, \beta) p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta)\ \mathrm{d}\mathbf{w} \\ &= \int p(t|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta)\ \mathrm{d}\mathbf{w} \end{aligned} \tag{69}$$

where the function is not relevant to the parameter $\boldsymbol{w}$. Since the (62) involves the convolution of two Gaussian distributions, the predictive distribution could be the form

$$p(t|\boldsymbol{x}, \mathbf{t}, \alpha, \beta) = \mathcal{N}(t|\mathbf{m}_N^T \boldsymbol{\Phi}(\boldsymbol{x}), \sigma_N^2(\boldsymbol{x})) \tag{70}$$

where

$$\sigma_N^2(\boldsymbol{x}) = \frac{1}{\beta} + \boldsymbol{\phi}^T \mathbf{S}_N \boldsymbol{\phi}(\boldsymbol{x}) \tag{71}$$

Note that, the first term represents the noise on the data, whereas the second term reflects the uncertainty associated with the parameters $\boldsymbol{w}$. And also, if the size of data set becomes larger, the second term goes to zero.

Predictive distribution

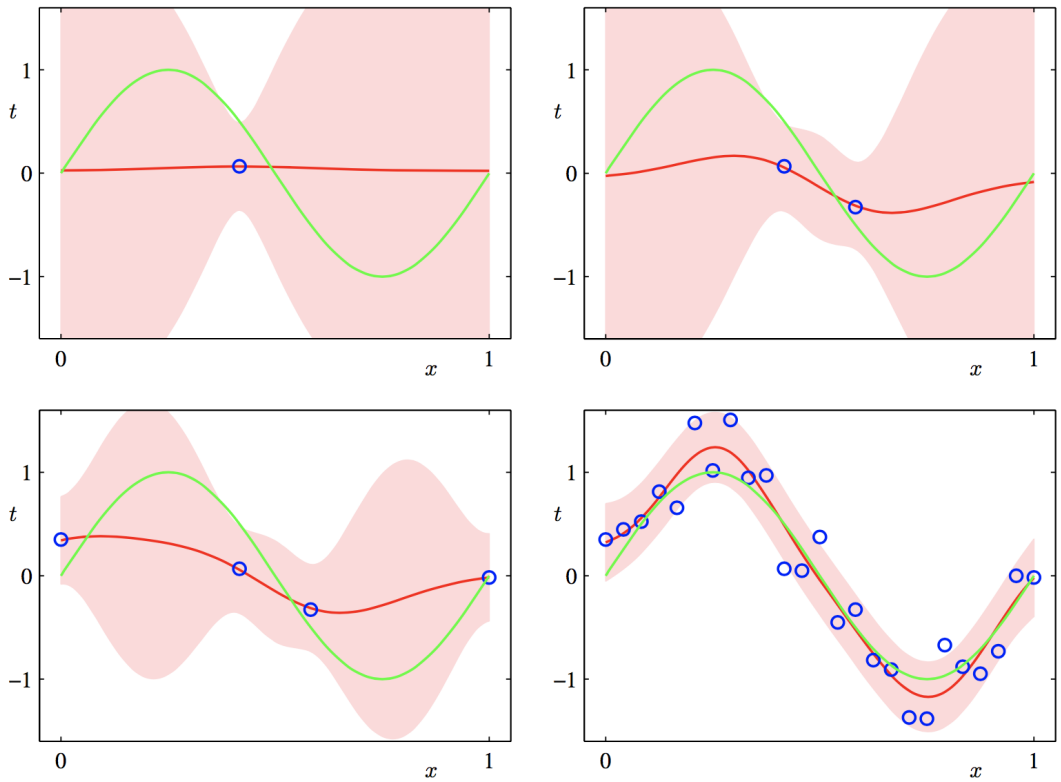

Figure 3: Examples of predictive distributions.

Model comparison

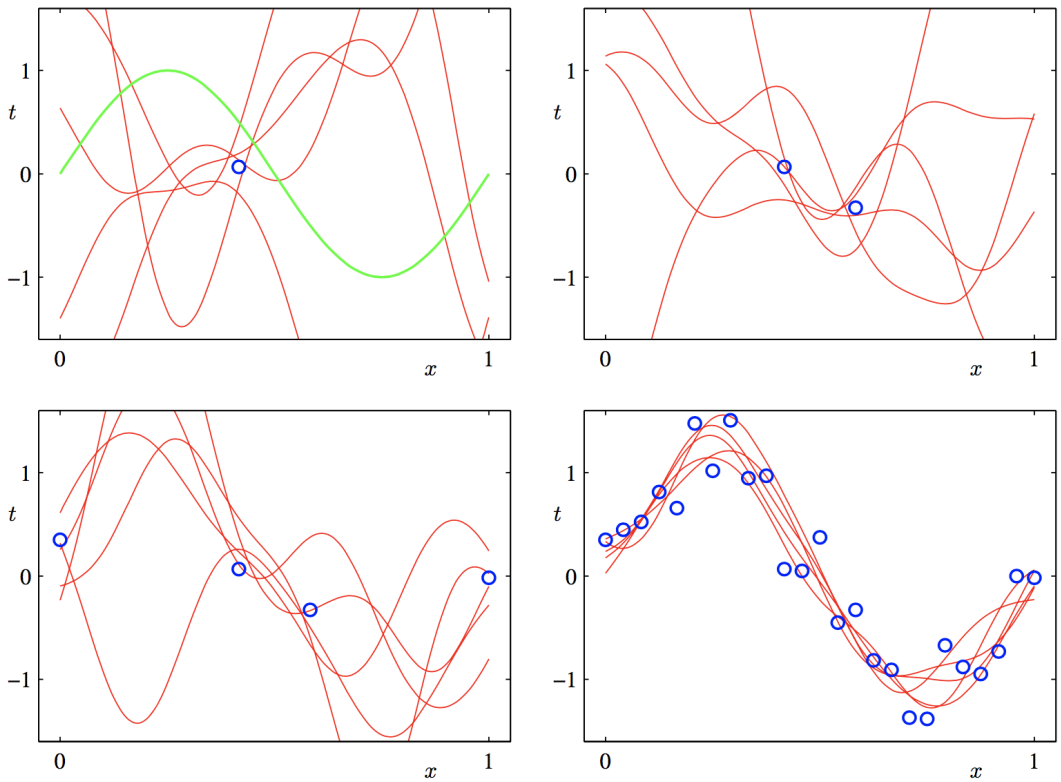

Figure 4: Plots of the function $y(x, \boldsymbol{w})$ from the posterior distributions.

Equivalent Kernel

We can see that the predictive mean as a specific model given by

$$y(\boldsymbol{x}, \mathbf{m}_N) = \mathbf{m}_N^T \boldsymbol{\phi}(\boldsymbol{x}) = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \mathbf{S}_N \boldsymbol{\Phi}^T \mathbf{t} = \sum_{n=1}^{N} \beta \boldsymbol{\phi}(\boldsymbol{x})^T \mathbf{S}_N \boldsymbol{\phi}(\boldsymbol{x}_n) t_n \tag{72}$$

where $\boldsymbol{w} = \mathbf{m}_N$. Then, the equation (65) may be interpreted as a linear combination of the target variables t_n from training set, so that we can write

$$y(\boldsymbol{x}, \mathbf{m}_N) = \sum_{n=1}^{N} k(\boldsymbol{x}, \boldsymbol{x}_n) t_n \tag{73}$$

where the function

$$k(\boldsymbol{x}, \boldsymbol{x}') = \beta \boldsymbol{\phi}(\boldsymbol{x})^T \mathbf{S}_N \boldsymbol{\phi}(\boldsymbol{x}') \tag{74}$$

is known as *the smoother matrix* or *the equivalent kernel*. And, also *the equivalent kernel* could be obtained by

$$\begin{aligned} cov[y(\boldsymbol{x}), y(\boldsymbol{x}')] &= cov[\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{w}, \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}')] \\ &= \boldsymbol{\phi}(\boldsymbol{x})^T \mathbf{S}_N \boldsymbol{\phi}(\boldsymbol{x}') \\ &= \beta^{-1} k(\boldsymbol{x}, \boldsymbol{x}') \end{aligned} \tag{75}$$

Note that, the predictive mean at nearby points will be highly correlated, wheras for more distant pairs of points the correlation will be smaller.

A Bayesian Viewpoint of The Model Complexity

What is model selection?

- Choosing the value of the hyperparameters.
- Choosing a model between alternative models.

How?

- Frequentist validates their model selection using validation set in their Variance-bias framework.
- Bayesian use the probabilistic framework to express the uncertainty in the choice of model.

Bayesian Model Comparison

Uncertainty of the model

Now, suppose we wish to compare a set of L models $\{\mathcal{M}_i\}$ and suppose that the data $\mathcal{D}$ is generated from one of these models, but we are uncertain which one.

$$data : \mathcal{D} \quad model : \{\mathcal{M}_i\} \forall i = \{1, \ldots, L\} \tag{76}$$

Prior distribution for models

Each model $\{\mathcal{M}_i\}$ refers to a probability distribution over the observed data $\mathcal{D}$. And our uncertainty over models would be expressed through a prior probability distribution

$$p(\mathcal{M}_i) \tag{77}$$

Posterior distribution for models

Given a training set $\mathcal{D}$, we then wish to evaluate the posterior distibution which is calculated by the Bayes' theorem given by

$$p(\mathcal{M}_i|\mathcal{D}) \propto p(\mathcal{M}_i)p(\mathcal{D}|\mathcal{M}_i) \tag{78}$$

The evidence term $p(\mathcal{D}|\mathcal{M}_i)$ also called the marginal likelihood function because the parameters $\boldsymbol{w}$ are marginalized out.

Predictive distribution

Predictive distribution

Once we know the posterior distribution over models, the predictive distribution is given by

$$p(t|\boldsymbol{x}, \mathcal{D}) = \sum_{i=1}^{L} p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D}) p(\mathcal{M}_i|\mathcal{D}) \tag{79}$$

The predictive distribution is marginalized over the space of models.

- $p(t|\boldsymbol{x}, \mathcal{M}_i, \mathcal{D})$: Predictive distributions of individual models
- $p(\mathcal{M}_i|\mathcal{D})$: Posterior probabilities of those models.

Model selection

A simple approximation to the predictive distribution marginalized over models is to use the single, most probable model alone to make predictions. This is known as *model selection*.

Evidence (Marginal Likelihood)

Evidence (Marginal Likelihood)
For a model governed by a set of parameters $\boldsymbol{w}$, the model evidence is given by

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w}|\mathcal{M}_i) \mathrm{d}\boldsymbol{w} \tag{80}$$

- $p(\mathcal{D}|\boldsymbol{w}, \mathcal{M}_i)$: probability of generating the data set $\mathcal{D}$ from a model $\mathcal{M}_i$
- $p(\boldsymbol{w}|\mathcal{M}_i)$: prior distribution for parameters $\boldsymbol{w}$

Simple Approximation to Evidence
Now, to get some insight into the model evidence, we approximate the integral. At first, let us assume some simplified situation as follows

- The posterior is picked around the most probable value w_{MAP}, with width $\Delta w_{\mathrm{posterior}}$
- The prior is flat with width $\Delta w_{\mathrm{prior}}$

Then, we get

$$p(\mathcal{D}|\mathcal{M}_i) = \int p(\mathcal{D}|\boldsymbol{w}, \mathcal{M}_i) p(\boldsymbol{w}|\mathcal{M}_i) \mathrm{d}\boldsymbol{w} \simeq p(\mathcal{D}|w_{\mathrm{MAP}}) \frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \tag{81}$$

Evidence (Marginal Likelihood)

and so taking logs we obtain

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|w_{\mathrm{WAP}}) + \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right) \tag{82}$$

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

$$\ln p(\mathcal{D}) \simeq \ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{WAP}}) + M \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right) \tag{83}$$

- **Fitness**: $\ln p(\mathcal{D}|\boldsymbol{w}_{\mathrm{WAP}})$
  The first term represents the fit to the data given by the most probable parameter values. As we increase the complexity of the model, it increases because a more complex model is better able to fit the data.
- **Penalty**: $M \ln \left(\frac{\Delta w_{\mathrm{posterior}}}{\Delta w_{\mathrm{prior}}} \right)$
  The second term represents how much the parameters tuned to the data. It penalizes the model according to its complexity. Since it has negative values, it decreases as we increase the complexity of the model.

Bayesian Model Comparison

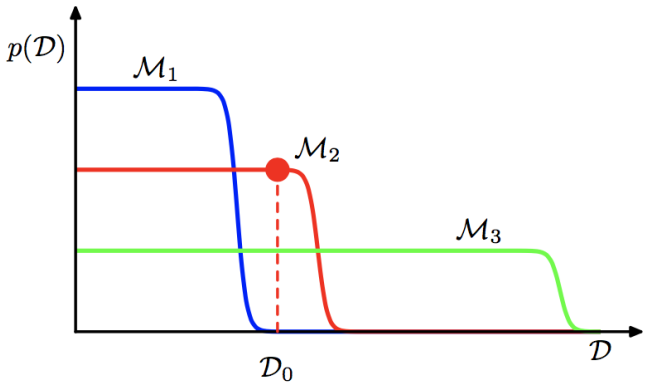

Figure 5: Distribution of possible data set of models.

How a specific data set $\mathcal{D}$ and its marginal likelihood $p(\mathcal{D})$ can favour models of intermediate complexity? Let's consider the case where we observe a specific data $\mathcal{D}_0$, then we may assume three models as follows

- Low complexity $\mathcal{M}_1$
  A simple model cannot fit the data well, whereas it assigns larger probability than more complex one.
- Intermediate complexity $\mathcal{M}_2$
  For specific data $\mathcal{D}_0$ will have the largest probability when we choose a model that has intermediate complexity. Simpler one can not fit to the data, whereas more complex one can fit to the data but it has smaller probabilty.
- High complexity $\mathcal{M}_3$
  A complex model spreads its predictive probability, so that it assigns relatively small probability.

Model Comparison and Bayes Factor

Let consdier two models $\mathcal{M}_1$ and $\mathcal{M}_2$, where the correct answer is the model $\mathcal{M}_1$. For a given finite data set, it is possible that the Bayes factor flavour incorrect answer which is given by

$$\frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} < 1 \tag{84}$$

However, if we average the Bayes factor over the distrubtion of data sets from true model

$$\int p(\mathcal{D}|\mathcal{M}_1) \ln \left(\frac{p(\mathcal{D}|\mathcal{M}_1)}{p(\mathcal{D}|\mathcal{M}_2)} \right) \mathrm{d}\mathcal{D} = D_{\mathrm{KL}}(\mathcal{M}_1||\mathcal{M}_2) \geq 0 \tag{85}$$

Thus on average the Bayes factor will always favour the correct model. For a given

The Evidence Approximation

The fully Bayesian approach
In the fully Bayesian approach, we would also specify a prior distribution over the hyperparameters, such as α and β, so that they also can be marginalized in the Bayesian framework.

$$p(t|\boldsymbol{x}, \mathbf{t}) = \iiint p(t|\boldsymbol{x}, \boldsymbol{w}, \beta) p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta) p(\alpha, \beta|\mathbf{t}) \, \mathrm{d}\boldsymbol{w} \mathrm{d}\alpha \mathrm{d}\beta \tag{86}$$

- **Our model**: $p(t|\boldsymbol{w}, \beta) = \mathcal{N}(t|y(\boldsymbol{x}, \boldsymbol{w}), \beta^{-1})$
- **posterior over parameters**: $p(\boldsymbol{w}|\mathbf{t}, \alpha, \beta) = \mathcal{N}(\boldsymbol{w}|\mathbf{m}_N, \mathbf{S}_N)$
- **posterior over hyperparameters**: $p(\alpha, \beta|\mathbf{t})$

Intractable solution and Evidence approximation
But the solution above is analytically untractable, so we need to evaluate the predictive model by approximating the model distribution. It is also called the *evidence approximation*

The Evidence Approximation

Assumptions for approximation

Let $p(\alpha, \beta|\mathbf{t})$ is sharply peaked around values $\hat{\alpha}$ and $\hat{\beta}$, then we get

$$p(t|\mathbf{t}) \simeq p(t|\mathbf{t}, \hat{\alpha}, \hat{\beta}) = \int p(t|\boldsymbol{w}, \hat{\beta}) p(\boldsymbol{w}|\mathbf{t}, \hat{\alpha}, \hat{\beta}) \mathrm{d}\boldsymbol{w} \tag{87}$$

Now, assuming that the prior is relatively flat, from the Bayes' theorem

$$p(\alpha, \beta|\mathbf{t}) \propto p(\mathbf{t}|\alpha, \beta) p(\alpha, \beta) \tag{88}$$

Evidence maximization

the values of $\hat{\alpha}$ and $\hat{\beta}$ are obtained by maximizing the marginal likelihood function which is given by

$$p(\mathbf{t}|\alpha, \beta) = \int p(\mathbf{t}|\boldsymbol{w}, \beta) p(\boldsymbol{w}|\alpha) \mathrm{d}\boldsymbol{w} \tag{89}$$

where

$$\begin{aligned} p(\mathbf{t}|\boldsymbol{w}, \beta) &= \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}_n), \beta^{-1}) \\ p(\boldsymbol{w}|\alpha) &= \mathcal{N}(\boldsymbol{w}|\mathbf{0}, \alpha^{-1}\mathbf{I}) \end{aligned} \tag{90}$$

We can re-estimate α and β analytically or algorithmically.

- Analytical way: evalute the evidence function, set its derivative equal to zero and then obtain re-estimation equations for α and β
- Alternatively, use the technique called the expected maximization algorithm.

The Evidence Approximation

Evaluate the evidence function analytically
The marginal likelihood is obtained by integrating out parameters

$$p(\mathbf{t}|\alpha,\beta) = \int p(\mathbf{t}|\boldsymbol{w},\beta)p(\boldsymbol{w}|\alpha)\mathrm{d}\boldsymbol{w} \tag{91}$$

and then we can write the evidence function, by (90), in the form

$$p(\mathbf{t}|\alpha,\beta) = \left(\frac{\beta}{2\pi}\right)^{N/2}\left(\frac{\alpha}{2\pi}\right)^{M/2}\int \exp\{-E(\boldsymbol{w})\}\mathrm{d}\boldsymbol{w} \tag{92}$$

where

$$\begin{aligned} E(\boldsymbol{w}) &= \beta E_D(\boldsymbol{w}) + \alpha E_W(\boldsymbol{w}) = \frac{\beta}{2}||\mathbf{t}-\boldsymbol{\Phi}\boldsymbol{w}||^2 + \frac{\alpha}{2}\boldsymbol{w}^T\boldsymbol{w} \\ &= E(\mathbf{m}_N) + \frac{1}{2}(\boldsymbol{w}-\mathbf{m}_N)^T\mathbf{A}(\boldsymbol{w}-\mathbf{m}_N) \end{aligned} \tag{93}$$

where

$$\mathbf{A} = \alpha\mathbf{I} + \beta\boldsymbol{\Phi}^T\boldsymbol{\Phi} \quad \mathbf{m}_N = \beta\mathbf{A}^{-1}\boldsymbol{\Phi}^T\mathbf{t} \tag{94}$$

The Evidence Approximation

Hence, we get

$$\int \exp\{-E(\boldsymbol{w})\}\,\mathrm{d}\boldsymbol{w} = \exp\{-E(\mathbf{m}_N)\}(2\pi)^{M/2}|\mathbf{A}|^{-1/2} \tag{95}$$

Taking logarithm to (91), using (95), the log of marginal likelihood is given by

$$\ln p(\mathbf{t}|\alpha,\beta) = \frac{M}{2}\ln\alpha + \frac{N}{2}\ln\beta - E(\mathbf{m}_N) - \frac{1}{2}\ln|\mathbf{A}| - \frac{N}{2}\ln(2\pi) \tag{96}$$

The Evidence Approximation

Set its derivative equal to zero (WRT α)
Defining the following eigenvector equation

$$(\beta \mathbf{\Phi}^T \mathbf{\Phi}) \boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i \tag{97}$$

then $\mathbf{A}$ has eigenvalues $\alpha + \lambda_i$ by

$$\mathbf{A}\boldsymbol{u}_i = (\alpha + \lambda_i)\boldsymbol{u}_i \tag{98}$$

Evaluate the derivative of the log of the marginal likelihood

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln p(\mathbf{t}|\alpha, \beta) &= \frac{\mathrm{d}}{\mathrm{d}\alpha}\left(\frac{M}{2} \ln \alpha\right) - \frac{\mathrm{d}}{\mathrm{d}\alpha}(\ln |\mathbf{A}|) - \frac{\mathrm{d}}{\mathrm{d}\alpha} E(\mathbf{m}_N) \\ &\simeq \frac{M}{2\alpha} - \frac{\mathrm{d}}{\mathrm{d}\alpha} \frac{1}{2} \ln \mathbf{A} - \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N \end{aligned} \tag{99}$$

Since

$$\frac{\mathrm{d}}{\mathrm{d}\alpha} \ln |\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\alpha} \ln \prod_{i=1}^{M} (\lambda_i + \alpha) = \sum_{i=1}^{M} \frac{1}{\lambda_i + \alpha} \tag{100}$$

then the stationary point is

$$0 = \frac{M}{2\alpha} - \frac{1}{2} \sum_{i=1}^{M} \frac{1}{\lambda_i + \alpha} - \frac{1}{2} \mathbf{m}_N^T \mathbf{m}_N \tag{101}$$

The Evidence Approximation

Rearrange the eqaution, then we get

$$\alpha \mathbf{m}_N^T \mathbf{m}_N = M - \alpha \sum_{i=1}^{M} \frac{1}{\lambda_i + \alpha} = \gamma \tag{102}$$

Since $M = \sum_{i=1}^{M} 1$, then

$$\alpha \mathbf{m}_N^T \mathbf{m}_N = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha} = \gamma \quad \text{or} \quad \alpha = \frac{\gamma}{\mathbf{m}_N^T \mathbf{m}_N} \tag{103}$$

Rearrange sequence

1. Set prior distribution $p(\boldsymbol{w})$ over parameters, get α, calculate γ
2. Evaluate posterior distribution $p(\boldsymbol{w}|\mathbf{t})$, get $\mathbf{m}_N$
3. Re-estimate α as $\gamma / \mathbf{m}_N^T \mathbf{m}_N$

The Evidence Approximation

Set its derivative equal to zero (WRT β)

$$\frac{\mathrm{d}}{\mathrm{d}\beta}\ln|\mathbf{A}| = \frac{\mathrm{d}}{\mathrm{d}\beta}\sum_i \ln(\lambda_i+\alpha) = \frac{1}{\beta}\sum_i \frac{\lambda_i}{\lambda_i+\alpha} = \frac{\gamma}{\beta} \tag{104}$$

Then, the stationary point is given by solving the following equation

$$0 = \frac{N}{2\beta} - \frac{1}{2}(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N)^T(\mathbf{t}-\mathbf{\Phi}\mathbf{m}_N) - \frac{\gamma}{2\beta} \tag{105}$$

Therefore, solving the equation, we get

$$\beta^{-1} = \frac{1}{N-\gamma}\sum_{i=1}^{N}\{t_n - \mathbf{m}_N^T\boldsymbol{\phi}(\boldsymbol{x}_n)\}^2 \tag{106}$$

An Elegant Interpretation of α and γ

Since, $\mathbf{\Phi}^T\mathbf{\Phi}$ is positive definitive, the eigenvector equation must have postivie eigenvalues $\lambda_i >= 0$. And then we get

$$0 \leq \frac{\lambda_i}{\lambda_i + \alpha} \leq 1 \tag{107}$$

And then, we may consider two cases of the value of λ_i

- $(\lambda_i \gg \alpha)$:
  The corresponding parameter w_i will be close to its maximum likelihood value. The value of w_i is tightly constrained by the data. Such parameters are called *well determined*
- $(\lambda_i \ll \alpha)$:
  The corresponding parameter w_i will be close to zero. The value of w_i is loosely constrained by the data.

Now, we can construct a feasible range of γ as follows

$$0 \leq \gamma = \sum_{i=1}^{M} \frac{\lambda_i}{\lambda_i + \alpha} \leq M \tag{108}$$

As the extension of the concept *well determined*, the quantity γ measures the effective total number of well determined parameters.