LECTURE 4: BAYESIAN INFERENCE

[1] Least Squares' Problems

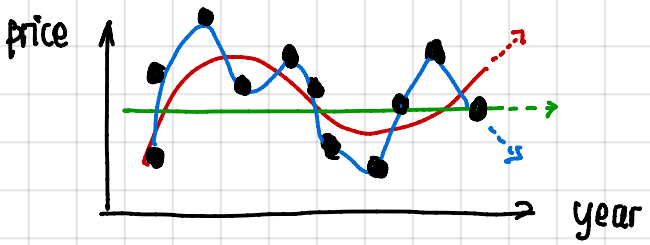
How to determine M?

↓

Model Selection problem:

- When M is too small:
  in the training stage, the error is high and the curve is under-fitting.
- When M is too large:
  in the training stage, the error is very low; but in the testing stage, the error is high (meaning, inaccurate inference). This is called overfitting.

⇓



Red line: the correct curve
Green line: an underfitting curve
Blue line: an overfitting curve

Validation process can address the overfitting problem, yet it assumes the distribution of its data represents that of the test data

↓

Which is not always true.

Least squares doesn't provide uncertainty / probability information

↓

Bayesian Inference

[2] Basic Probability Theory

Main task of Machine Learning : to make inferences

⇓

Type of Inferences

Inductive:

to reach probable conclusions.
All needed information is unavailable or unknown, causing uncertainty in the conclusions.

Deductive :

to reach logical conclusions deterministically : All information that can lead to the correct conclusion is available.

↓

Probability & Statistics are required → Statistical Machine Learning (as opposed to Logic machine learning)

Basic properties:

1. $0 \leq p(x) \leq 1$
2. $\int p(x)\,dx = 1$

↓

$$\sum_x p(x) = 1$$

e.g. $p(x = \text{Head}) + p(x = \text{Tail}) = 1$

Basic rules:

Product rule :

- independent variables:
$$p(a,b) = p(a)\,p(b)$$
- dependent variables:
$$p(a,b) = p(a|b)\,p(b) = p(b|a)\,p(a)$$

↓ see page #5 for some example

Sum rule :

- dependent variables:
$$p(a) = \sum_b p(a,b)$$

↓

Marginalization

Bayes' theorem :

$$p(x,d) = p(x|d)\,p(d)$$

$$p(x|d) = \frac{p(x,d)}{p(d)} = \frac{p(d|x)\,p(x)}{p(d)}$$

$$p(d|x) = \frac{p(x,d)}{p(x)}$$

[3] Bayesian Inference

Bayes' theorem:

$$p(x|d) = \frac{p(d|x)\,p(x)}{p(d)}$$

where: x = the random variable (or latent/hidden variable)
d = the observed data

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

$$p(x|d) = \frac{p(d|x)\,p(x)}{p(d)} = \frac{p(d|x)\,p(x)}{\int p(x,d)\,dx} = \frac{p(d|x)\,p(x)}{\int p(d|x)\,p(x)\,dx}$$

We will examine all these terms later

[4] Probability : Meaning

Frequentist:

We need to repeat the events many time & measure the mean and the variance.

Bayesian:

We don't need to repeat the event (we don't need to have event happens beforehand), as we can use the Bayes' formula the get the mean & variance.

[5] Bayesian-based Inference Methods

(1) MLE [Maximum Likelihood Estimation]

(2) MAP [Maximum A-Posteriori]

(3) Full Bayesian Inference (discussed further later)

Likelihood: $p(d|x)$

MLE: $x^* = \underset{\{x\}}{\arg\max}\ p(d|x)$

Posterior: $p(x|d)$

MAP: $x^* = \underset{\{x\}}{\arg\max}\ p(x|d)$

$= \underset{\{x\}}{\arg\max}\ p(d|x)\,p(x)$

Intuitive Meaning of MLE

Basic idea: To try every possible value of x and find the most probable one.

Let: d = a set of symptomps; $\{x\}$ = a set of possible illness

e.g.: $\{x\}$ = {cold, cancer, malaria}

$p(d|x) = p(\text{symptomps} \mid x = \text{malaria})$

Meaning: if we assume x (the illness) is malaria, how likely it can explain the observed symptomps.

If $p(\text{symptomps} \mid x = \text{malaria}) > p(\text{symptomps} \mid x = \text{cancer})$,

it means the probability of x to be malaria is more likely than that of to be cancer.

Note 1 : Example of Product Rule

1. Independent variables :
For a coin tossed twice, what is the probability of having (H, H)?

$$p(x=H, y=H) = p(x=H)\, p(y=H)$$
$$= \frac{1}{2} \cdot \frac{1}{2} = \frac{1}{4}$$

2. Dependent variables:
Given 2 markers with different colors (Red and Black), what is the probability of having B and then R.

The space of outcomes of taking two markers sequentially:

B - R
R - B
} thus having B-R is $\frac{1}{2}$

Mathematically:

$$p(x=B, y=R) = \boxed{p(y=R \mid x=B)}\ p(x=B)$$
$$= 1 \cdot \frac{1}{2} = \frac{1}{2}$$

$p(x=B)$ → $\frac{1}{2}$ since both markers are hidden, thus the probability of getting B is $\frac{1}{2}$

$p(y=R \mid B) = 1$, because we're given that the first pick was B, and know the probability of the last marker to be R must be 1.

Another example:

Given 3 markers (R, G, B), what is the probability of having (B and then G) we pick 2 randomly.

Space of outcomes:

R-G	G-B	B-R
R-B	G-R	B-G

} Thus: $p(y=G, x=B) = \frac{1}{6}$

Mathematically:

$$p(y=G, x=B) = \boxed{p(y=G \mid x=B)}\ p(x=B)$$
$$= \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

$p(x=B) = \frac{1}{3}$ → if we have 3 hidden markers, the chance of the first random pick is B, is $\frac{1}{3}$

$p(y=G \mid x=B)$ → Given that the first pick is B, there are two hidden markers, and the chance of picking G is $1/2$.

MLE FOR REGRESSION

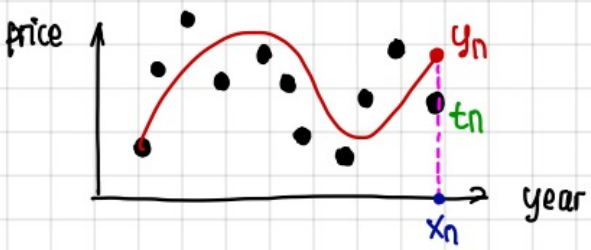
Motivation:

(1) There are 2 main drawbacks of least squares: Model selection problem (we don't know how to set M), and no uncertainty information (probability) on the decision, making us unable to know how good the prediction is, particularly in the testing stage.

(2) Using MLE, we hope to solve these 2 drawbacks, but can't it?

↓↓

Step 1: Likelihood data modeling:



Real data always suffer from noise. Hence, in the training, fitting to them perfectly will cause an overfitting problem.

→ $\bar{x}$ R.V ↓↓

Implying: $\bar{y}$ (the estimated curve) must represent the underlying pattern of the data ($\bar{t}$), and exclude noise as much as possible.

↳ observe

Correlation of $\bar{y}$ & $\bar{t}$ without considering noise: already know → d.

$$t_n = y_n = \bar{w}^T \bar{x}_n$$

With noise, the correlation becomes: sampled from

$$t_n = y_n + \varepsilon \quad ; \quad \varepsilon \sim G(0, \sigma^2)$$

noise

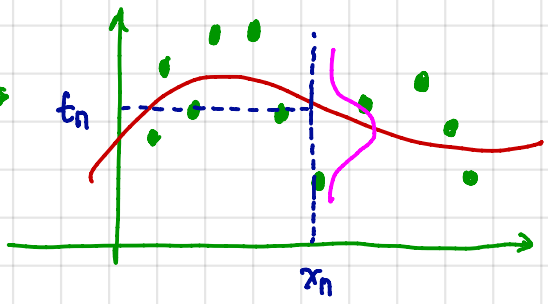
Thus:

$$p(t_n \mid x_n, \bar{w}, \sigma^2) = G(t_n ; y_n, \sigma^2)$$

$$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_n - t_n)^2}{2\sigma^2}\right)$$

$$\log G(t_n; y_n, \sigma^2) = \log \frac{1}{\sqrt{2\pi}} + \log \frac{1}{\sigma} + \left(\frac{-(y_n - t_n)^2}{2\sigma^2}\right) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \sim$$

$$p(t_n \mid x_n, \bar{w}, \sigma^2) = G(t_n; \bar{w}^T \bar{x}_n, \sigma^2) \rightarrow$$



Assuming the samples are independent to each other, for all samples, the likelihood can be expressed as:

$$p(\bar{t} \mid \bar{x}, \bar{w}, \sigma^2) = \prod_n G(t_n; y(x_n, \bar{w}), \sigma^2)$$

e.g.: $p(t_1, t_2) = p(t_1)\, p(t_2)$

Step 2 : Error function

MLE (maximizing likelihood) means:

$$\{\bar{w}, \sigma^2\}^* = \underset{\{\bar{w}\}\{\sigma^2\}}{\text{argmax}}\ p(\bar{t} \mid \bar{x}, \bar{w}, \sigma^2) \qquad y(x_n, \bar{w}) = \bar{w}^T \bar{x}_n$$

$$= \underset{\{\bar{w}\}\{\sigma^2\}}{\text{argmax}} \sum_n \log G(t_n; y_n, \sigma^2)$$

minimization is more common than maximization →

$$= \underset{\{\bar{w}\}\{\sigma^2\}}{\text{argmin}} \sum_n -\log G(t_n; \bar{w}^T \bar{x}_n, \sigma^2)$$

$$= \underset{\{\bar{w}\}\{\sigma^2\}}{\text{argmin}} \underbrace{\sum_n \log\sqrt{2\pi} + \sum_n \frac{1}{2}\log\sigma^2 + \sum_n \frac{(\bar{w}^T \bar{x}_n - t_n)^2}{2\sigma^2}}_{E(\bar{w}, \sigma^2)}$$

Step 3: Optimization

$2(\bar{w}^T\bar{x}_n - t_n)\cdot\bar{x}_n$ (1×M, M×1, 1×1, M×1)

For $\bar{w}$: $\frac{\partial E}{\partial \bar{w}} = \frac{\partial}{\partial \bar{w}}\left(\sum_n \frac{(\bar{w}^T\bar{x}_n - t_n)^2}{2\sigma^2}\right) = 0$

$$= \frac{1}{2\sigma^2}\sum_n \frac{\partial}{\partial \bar{w}}(\bar{w}^T\bar{x}_n - t_n)^2$$

$$= \frac{1}{\sigma^2}\sum_n (\underbrace{\bar{w}^T\bar{x}_n}_{1\times 1} - t_n)\underbrace{\bar{x}_n}_{(M+1)\times 1} = 0$$

Recall that: $\mathbb{X}$ = a $N \times (M+1)$ matrix

$$= \underbrace{\mathbb{X}^T\mathbb{X}}_{(M+1)\times(M+1)}\underbrace{\bar{w}}_{(M+1)\times 1} - \underbrace{\mathbb{X}^T}_{(M+1)\times N}\underbrace{\bar{t}}_{N\times 1} = 0$$

$(M+1)\times 1$ $\quad (M+1)\times 1$

$\mathbb{X}^T\mathbb{X}\bar{w} = \mathbb{X}^T\bar{t}$

$\bar{w} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T\bar{t}$

$$\boxed{\bar{w} = \underbrace{(\mathbb{X}^T\mathbb{X})^{-1}}_{M\times N}\underbrace{\mathbb{X}^T}_{N\times 1}\bar{t} = 0 = \underbrace{\mathbb{X}^\dagger}_{M\times N}\underbrace{\bar{t}}_{N\times 1}}$$

Exactly the same as least squares!

» Unlike least squares, MLE provides an opportunity to know the probability of our estimation with respect to the ground truths, $\bar{t}$:

$p(\bar{t} \mid \bar{x}, \bar{w}, \sigma^2)$ → We can know how good our estimated $\bar{w}$

↓↓

However, to get the probability value, we need to know σ^2; otherwise, we can't compute $p(\bar{t} \mid \bar{x}, \bar{w}, \sigma^2)$.

For σ^2: $\frac{\partial E}{\partial \sigma^2} = 0$ (Why is it true? see a note in the last page)

$$\frac{\partial}{\partial \sigma^2}\left(\sum_n \log\sqrt{2\pi} + \sum_n \frac{1}{2}\log\sigma^2 + \sum_n \frac{(\bar{w}^T x_n - t_n)^2}{2\sigma^2}\right) = 0$$

$\frac{1}{2}\log\sigma^2$ → $\frac{d\log\sigma^2}{d\sigma^2} = \frac{d\log u}{du} = \frac{1}{u} = \sigma^{-2}$

$$\frac{N}{2}\sigma^{-2} - \frac{1}{2}\sigma^{-4}\sum_n (\bar{w}^T\bar{x}_n - t_n)^2 = 0$$

Hence:

$$\sigma^2 = \frac{1}{N}\sum_n (\bar{w}^T\bar{x}_n - t_n)^2$$

↳ this is known

↓↓

By having σ^2, we can compute the uncertainty / probability:

$$p(\bar{t}\,|\,\bar{x}, \bar{w}, \sigma^2) = \prod_n G(t_n; y_n, \sigma^2)$$

$$= \prod_n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_n - t_n)^2}{2\sigma^2}\right)$$

$$\bar{w}^T\bar{x}_n = \sum_{m=0}^{M} w_m x_n^m$$

↓↓

Unfortunately, σ^2 depends on M:
If M is large, σ^2 is very small, and if M is small, σ^2 is large.

↓↓

Hence, σ^2, which determines the uncertainty / probability, depends significantly on M, which is set manually & can be wrong. Therefore, the computed σ^2 is meaningless.

↓↓

Implying that we cannot obtain uncertainty using MLE.

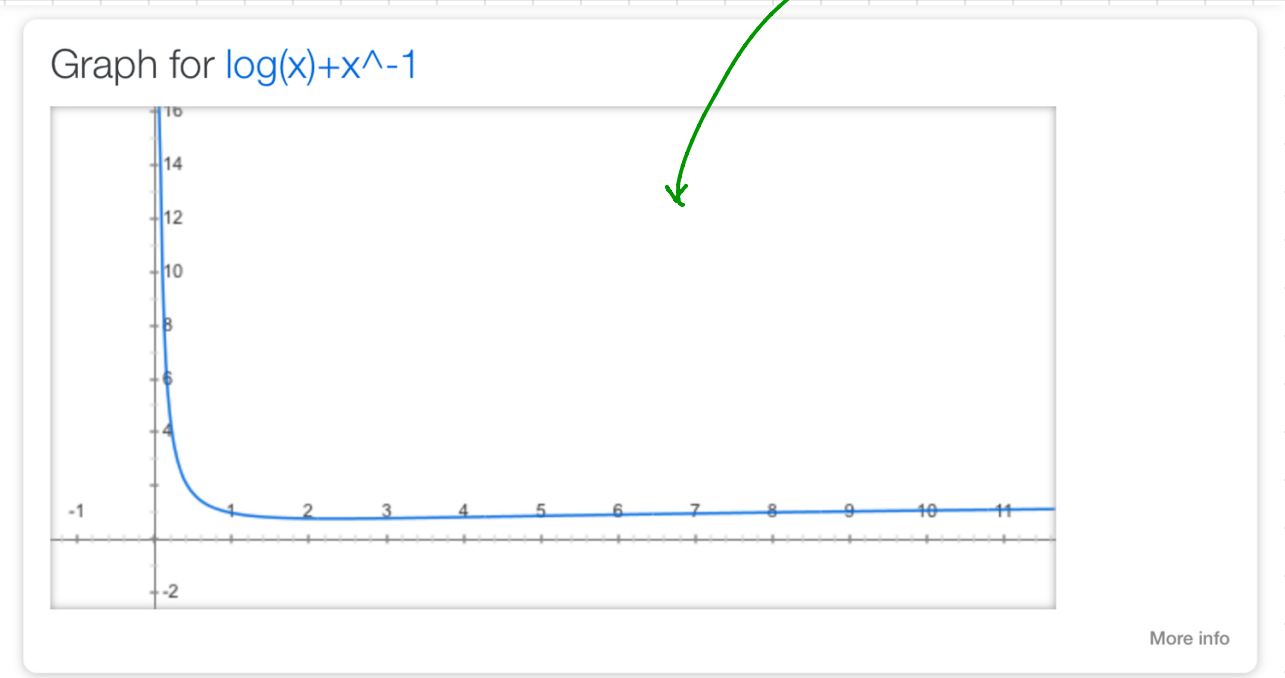
Also note: Estimating $\bar{w}$ is also independent of σ^2. Hence, there is no use of estimating / knowing σ^2.

Additional Notes:

Q: Why $\frac{\partial E}{\partial \sigma^2} = 0$ holds?

where: $E(\bar{w}, \sigma^2) = \sum_n \log\sqrt{2\pi} + \sum_n \frac{1}{2}\log\sigma^2 + \sum_n \frac{(\bar{w}^T x_n - t_n)^2}{2\sigma^2}$

A: The easiest way is to plot the function & see if there is only a single stationary point:



convex function.

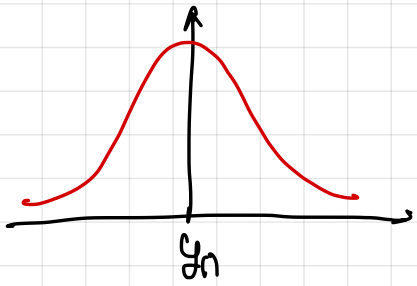
To prove mathematically that there is a single stationary point is more difficult, and beyond our discussion.

Q: Then, how do we know if $\frac{\partial E}{\partial x} = 0$ is correct?

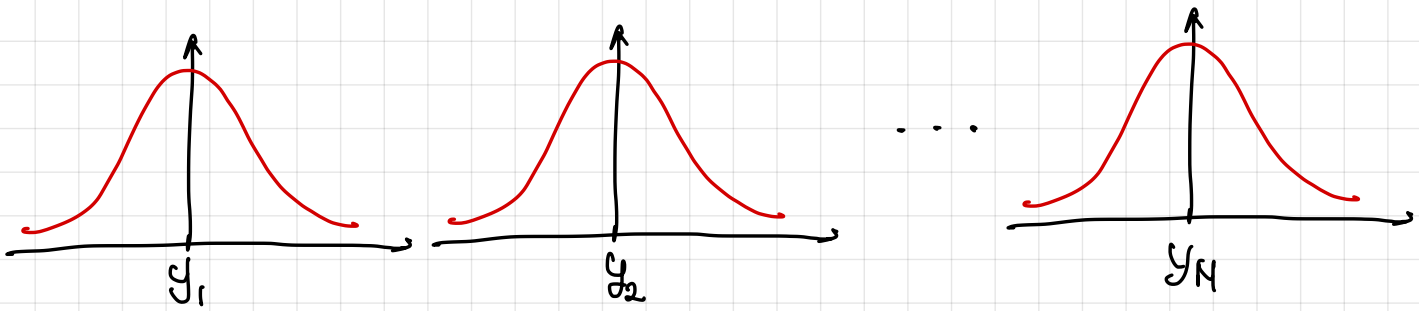
A: Just do it and if the derivation provide a closed-form solution, then the equation of E must be convex, with a single stationary point.

Likelihood Data Modeling:

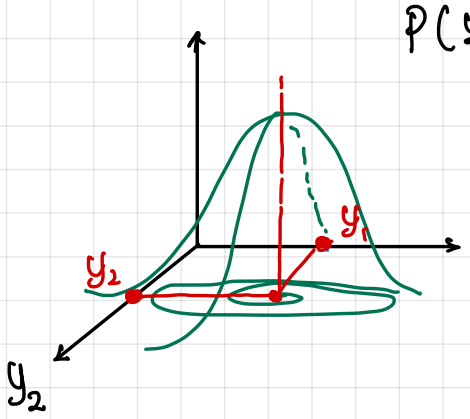
1. One sample (t_n):



2. More than one (independent):



3. Two samples (dependent):



$$P(y_1, y_2 \mid t_1, t_2) = G(t_1, t_2; y_1, y_2, \Sigma)$$

$$= G(\underset{N\times 1}{\bar{t}}; \underset{N\times 1}{\bar{y}}, \underset{N\times N}{\Sigma})$$

$$= \frac{1}{(2\pi)^{k/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left(-\underbrace{(\bar{y}-\bar{t})^T}_{1\times N} \underbrace{\Sigma^{-1}}_{N\times N} \underbrace{(\bar{y}-\bar{t})}_{N\times 1}\right)$$

$$\downarrow\downarrow$$

$$\underset{M\times 1}{\frac{\partial E}{\partial \bar{w}}} = -\frac{\partial}{\partial \bar{w}}\left((\bar{y}-\bar{t})^T \Sigma^{-1} (\bar{y}-\bar{t})\right) \quad ; \quad \underset{N\times 1}{\bar{y}} = \underset{N\times M}{X}\, \underset{M\times 1}{\bar{w}}$$

$$= -\underbrace{(\bar{y}-\bar{t})}_{N\times 1} \underset{N\times N}{\Sigma^{-1}} \underset{N\times M}{X} = -X^T \Sigma^{-1} (\bar{y}-\bar{t}) = 0$$

$$X^T \Sigma^{-1} (X\bar{w}) = X^T \Sigma^{-1} \bar{t}$$

$$\underset{M\times 1}{\bar{w}} = \underbrace{\left(X^T \Sigma^{-1} X\right)^{-1}}_{M\times M} \underbrace{X^T}_{M\times N} \underbrace{\Sigma^{-1}}_{N\times N} \underbrace{\bar{t}}_{N\times 1}$$

MAP FOR REGRESSION

Motivation:

» MLE cannot solve the two main problems we have:

(1) Model selection problem

(2) Uncertainty information (probability of decision)

Also:

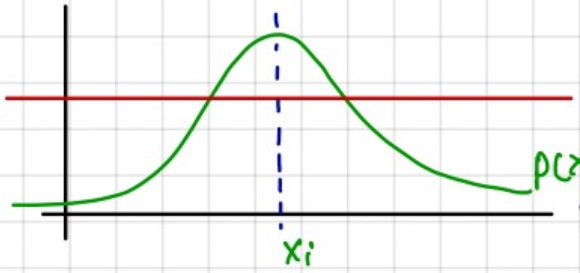
MLE has another main drawback: There is no constraint or prior information on x. In other words, all candidates $\{x\}$ have the same treatment.

MAP solves the MLE's drawback by incorporating prior information on x:

$$p(x|d) = \frac{p(d|x)\ p(x)}{p(d)}$$

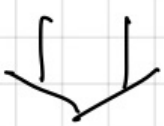
$$x^* = \underset{\{x\}}{\arg\max}\ p(x|d) = \underset{\{x\}}{\arg\max}\ p(d|x)\ p(x)$$ → since $p(d)$ is constant w.r.t. x.

e.g.



$p(x)$ → MAP becomes MLE

$p(x)$ → different treatments for different values of x.



Q: Can MAP solve the model selection problem and the uncertainty problem?

[2] MAP : Regression

(1) Posteriori Data Modeling :

Input : t_n , x_n

Output : $y_n = \sum_m^{} w_m x_n^m = \bar{w}^T \bar{x}_n$ } one sample

Posterior :

$$p(y_n | t_n) = p(\bar{w}, \bar{x}_n | t_n)$$

$$= p(t_n, \bar{w}, x_n) / p(t_n)$$

$$= \frac{p(t_n | \bar{w}, x_n)}{p(t_n)} p(\bar{w}, x_n)$$

Likelihood

dependent : $p(\bar{w} | x_n)\, p(x_n)$

independent : $p(\bar{w})\, p(x_n)$

* $p(x_n) = ?$ For our case, it's difficult to find any prior on x_n. Hence, let $p(x_n) = 1$

* $p(\bar{w} | x_n) = ?$ For our case, it's also difficult to define any function of $\bar{w}$ given any values of x_n. As $\bar{w}$ depends on both x_n & t_n.

* $p(\bar{w}) = ?$

The likelihood prefers M to be high, so that $p(t_n | y_n)$ can be high; but we know high M means overfitting. Hence, we should encourage $\bar{w}$ (i.e., $w_0, w_1, \ldots, w_{M-1}$) to be zero (or small) : $\min \left(\sum_{m=0}^{M-1} |w_m| \right)$

$$p(\bar{w}) = G(\bar{w}; 0, \alpha^{-1} I)$$

Posterior : $$p(y_n | t_n) = G(t_n; \bar{w}^T \bar{x}_n, \beta^{-1})\, G(\bar{w}; 0, \alpha^{-1} I)$$

where : $\beta^{-1} = \sigma^2$

$= \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(y_n - t_n)^2}{2\sigma^2}\right)$

Posterior for all samples:

$$p(\underset{N\times 1}{\bar{y}} \mid \underset{N\times 1}{\bar{t}}) = \frac{\prod_n^N G(t_n ; \bar{w}^T \bar{x}_n , \beta^{-1}) \; G(\underset{M\times 1}{\bar{w}} ; \underset{M\times 1}{0} , \underset{M\times M}{\alpha^{-1} I})}{p(\bar{t})}$$

$\alpha = \frac{1}{\sigma^2}$

$$p(\bar{w} ; 0, \alpha^{-1} I) = \left(\frac{\alpha}{2\pi}\right)^{\frac{M+1}{2}} \exp\left(-\frac{\alpha}{2} \bar{w}^T w\right)$$

(2) Error Function: (Max. A Posteriori)

$$\bar{w}^* = \underset{\{\bar{w}\}}{\text{argmax}} \; p(\bar{y} \mid \bar{t})$$

$$= \underset{\{\bar{w}\}}{\text{argmax}} \prod_n^N G(t_n ; y_n , \beta^{-1}) \; G(\bar{w} ; 0, \alpha^{-1} I)$$

$$= \underset{\{\bar{w}\}}{\text{argmax}} \sum_n^N \left[\log \frac{1}{\sqrt{2\pi\beta^{-1}}} - \frac{\beta}{2} (\bar{w}^T \bar{x}_n - t_n)^2 \right] + \log \left(\frac{\alpha}{2\pi}\right)^{M/2} - \frac{\alpha}{2} \bar{w}^T \bar{w}$$

$$= \underset{\{\bar{w}\}}{\text{argmax}} \; \underbrace{-\frac{N}{2}\log 2\pi + \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_n (\bar{w}^T \bar{x}_n - t_n)^2 - \frac{M}{2}\log 2\pi + \frac{M}{2}\log\alpha - \frac{\alpha}{2}\bar{w}^T\bar{w}}_{-E(\bar{w})}$$

$$= \underset{\{\bar{w}\}}{\text{argmin}} \; E(\bar{w})$$

Minimization: $\frac{\partial E(\bar{w})}{\partial \bar{w}} = 0$

$$\frac{\partial}{\partial \bar{w}} \left[\frac{\beta}{2} \sum_n (\bar{w}^T \bar{x}_n - t_n)^2 + \frac{\alpha}{2} \bar{w}^T \bar{w} \right] = 0$$

$$\beta \sum_n (\bar{w}^T \bar{x}_n - t_n) \bar{x}_n + \alpha \bar{w} = 0$$

$\bar{w}^T \bar{x}_n$: 1×1; $\bar{x}_n$: $(M+1) \times 1$; $\bar{w}$: $(M+1) \times 1$

$$X^T X \bar{w} - X^T \bar{t} + \frac{\alpha}{\beta} \bar{w} = 0$$

$$\left(X^T X + \frac{\alpha}{\beta} I \right) \bar{w} = X^T \bar{t} \longrightarrow \boxed{\bar{w} = \left(X^T X + \frac{\alpha}{\beta} I \right)^{-1} X^T \bar{t}}$$

Unlike MLE, in MAP $\bar{w}$ depends on the noise level β, which is better.

MAP (the prior) eases the overfitting problem, but doesn't solve the problem completely.

$$\beta (X^T X \bar{w} - X^T \bar{t}) + \alpha \bar{w} = 0$$

$$\beta X^T X \bar{w} - \beta X^T \bar{t} + \alpha \bar{w} = 0$$

$$\beta X^T X \bar{w} + \alpha \bar{w} = \beta X^T \bar{t}$$

$$(\beta X^T X + \alpha I) \bar{w} = \beta X^T \bar{t}$$

$$\boxed{\bar{w} = \beta (\beta X^T X + \alpha I)^{-1} X^T \bar{t}}$$

Q: Why don't we estimate β? Isn't β useful to indicate the uncertainty information (since, unlike in MLE, in MAP a larger M doesn't always lead to the overfitting problem?

A: The outcome of $p(t|x)\,p(x)$ is not a probability value!

MLE: $\{\bar{w}, \beta\}^* = \underset{\{\bar{w}\}\{\beta\}}{\operatorname{argmax}} \underbrace{p(\bar{t}|\bar{w}, \bar{x}, \beta)}_{\text{a probability function, where } \beta \text{ indicates the uncertainty}} = \underset{\{w\}\{\beta\}}{\operatorname{argmax}} \prod_n G(t_n; \bar{w}^T x_n, \beta^{-1})$

MAP: $\{\bar{w}\}^* = \underset{\{w\}}{\operatorname{argmax}} \underbrace{p(\bar{t}|\bar{w}, \bar{x}, \beta)\, p(\bar{w})}$

① the multiplication of these two probability functions is not likely to be a probability function.

② β only indicates the uncertainty of $p(\bar{t}|\bar{w}, \bar{x}, \beta)$ but not for $p(\bar{w})$.

Thus, computing β to determine the confidence on $\bar{w}$ is not applicable.