Introduction to Machine Learning

Maximum A Posteriori (MAP)

Andres Mendez-Vazquez

February 16, 2023

Outline

Cinvestav

Outline

Cinvestav

Big Problem

We have that we depend on the distribution we choose

- When using the Likelihood... Can do better?

Actually, Yes

- Something that comes from the Bayesian idea...

Big Problem

We have that we depend on the distribution we choose

- When using the Likelihood... Can do better?

Actually, Yes

- Something that comes from the Bayesian idea...

Introduction

We go back to the Bayesian Rule

$$p\left(\Theta|\mathcal{X}\right) = \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{p\left(\mathcal{X}\right)} \quad (1)$$

We now seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p\left(\Theta|\mathcal{X}\right)$

Introduction

We go back to the Bayesian Rule

$$p(\Theta|\mathcal{X}) = \frac{p(\mathcal{X}|\Theta)\, p(\Theta)}{p(\mathcal{X})} \quad (1)$$

We now seek that value for Θ, called $\widehat{\Theta}_{MAP}$

It allows to maximize the posterior $p(\Theta|\mathcal{X})$

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)$$

$$= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}$$

$$\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)$$

$$= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i\in\mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$
\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right)\\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)}{P\left(\mathcal{X}\right)}\\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right)p\left(\Theta\right)\\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right)p\left(\Theta\right)
\end{aligned}
$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

Development of the solution

We look to maximize $\widehat{\Theta}_{MAP}$

$$\begin{aligned}
\widehat{\Theta}_{MAP} &= \underset{\Theta}{\text{argmax}}\, p\left(\Theta|\mathcal{X}\right) \\
&= \underset{\Theta}{\text{argmax}} \frac{p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right)}{P\left(\mathcal{X}\right)} \\
&\approx \underset{\Theta}{\text{argmax}}\, p\left(\mathcal{X}|\Theta\right) p\left(\Theta\right) \\
&= \underset{\Theta}{\text{argmax}} \prod_{x_i \in \mathcal{X}} p\left(x_i|\Theta\right) p\left(\Theta\right)
\end{aligned}$$

$P\left(\mathcal{X}\right)$ can be removed because it has no functional relation with Θ.

We can make this easier

Use logarithms

$$\widehat{\Theta}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i|\Theta\right) + \log p\left(\Theta\right) \right] \quad (2)$$

Outline

Cinvestav

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

Cinvestav

What Does the MAP Estimate Get?

Something Notable

The MAP estimate allows us to inject into the estimation calculation our prior beliefs regarding the parameters values in Θ.

For example

Let's conduct N independent trials of the following Bernoulli experiment with q parameter:

- We will ask each individual we run into in the hallway whether they will vote PRI or PAN in the next presidential election.

With probability q to vote PRI

Where the values of x_i is either PRI or PAN.

Cinvestav

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ =& \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ & \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ =& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log\ p\left(\mathcal{X}|q\right) &= \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\ &= \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\ &\quad \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\ &= n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned}
\log\ p\left(\mathcal{X}|q\right) =& \sum_{i=1}^{N} \log\ p\left(x_i|q\right) \\
=& \sum_{i} \log\ p\left(x_i = PRI|q\right) + ... \\
& \sum_{i} \log\ p\left(x_i = PAN|1-q\right) \\
=& n_{PRI} \log\left(q\right) + \left(N - n_{PRI}\right) \log\left(1-q\right)
\end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

First the Maximum Likelihood Estimate

Samples

$$\mathcal{X} = \left\{ x_i = \begin{cases} PAN \\ PRI \end{cases} \quad i = 1, ..., N \right\} \tag{3}$$

The log likelihood function

$$\begin{aligned} \log\ p(\mathcal{X}|q) &= \sum_{i=1}^{N} \log\ p(x_i|q) \\ &= \sum_{i} \log\ p(x_i = PRI|q) + ... \\ & \sum_{i} \log\ p(x_i = PAN|1-q) \\ &= n_{PRI} \log(q) + (N - n_{PRI}) \log(1-q) \end{aligned}$$

Where n_{PRI} are the numbers of individuals who are planning to vote PRI this fall

We use our classic tricks

By setting

$$\mathcal{L} = \log \ p\left(\mathcal{X}|q\right) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} = 0 \tag{6}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log\ p(\mathcal{X}|q) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} = 0 \tag{6}$$

We use our classic tricks

By setting

$$\mathcal{L} = \log \; p\left(\mathcal{X}|q\right) \tag{4}$$

We have that

$$\frac{\partial \mathcal{L}}{\partial q} = 0 \tag{5}$$

Thus

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} = 0 \tag{6}$$

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \tag{7}$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Final Solution of ML

We get

$$\widehat{q}_{PRI} = \frac{n_{PRI}}{N} \quad (7)$$

Thus

If we say that $N = 20$ and if 12 are going to vote PRI, we get $\widehat{q}_{PRI} = 0.6$.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Building the MAP estimate

Obviously we need a prior belief distribution

We have the following constraints:

- The prior for q must be zero outside the $[0, 1]$ interval.
- Within the $[0, 1]$ interval, we are free to specify our beliefs in any way we wish.
- In most cases, we would want to choose a distribution for the prior beliefs that peaks somewhere in the $[0, 1]$ interval.

We assume the following

- The state of Colima has traditionally voted PRI in presidential elections.
- However, on account of the prevailing economic conditions, the voters are more likely to vote PAN in the election in question.

Cinvestav

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha - 1} (1 - q)^{\beta - 1}. \quad (8)$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \quad (9)$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha - 1} (1 - q)^{\beta - 1}. \tag{8}$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \tag{9}$$

What prior distribution can we use?

We could use a Beta distribution being parametrized by two values α and β

$$p(q) = \frac{1}{B(\alpha, \beta)} q^{\alpha - 1} (1 - q)^{\beta - 1}. \quad (8)$$

Where

We have $B(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$ is the beta function where Γ is the generalization of the notion of factorial in the case of the real numbers.

Properties

When both the $\alpha, \beta > 0$ then the beta distribution has its mode (Maximum value) at

$$\frac{\alpha - 1}{\alpha + \beta - 2}. \quad (9)$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

We then do the following

We do the following

We can choose $\alpha = \beta$ so the beta prior peaks at 0.5.

As a further expression of our belief

We make the following choice $\alpha = \beta = 5$.

Why? Look at the variance of the beta distribution

$$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}. \tag{10}$$

Cinvestav

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var\left(q\right) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Thus, we have the following nice properties

We have a variance with $\alpha = \beta = 5$

$Var\left(q\right) \approx 0.025$

Thus, the standard deviation

$sd \approx 0.16$ which is a nice dispersion at the peak point!!!

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{11}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{12}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1}\right) \tag{13}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i | q\right) + \log p\left(q\right) \right] \tag{11}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{12}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1}\right) \tag{13}$$

Now, our MAP estimate for $\widehat{p}_{MAP}$...

We have then

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[\sum_{x_i \in \mathcal{X}} \log p\left(x_i|q\right) + \log p\left(q\right) \right] \tag{11}$$

Plugging back the ML

$$\widehat{p}_{MAP} = \underset{\Theta}{\text{argmax}} \left[n_{PRI} \log q + \left(N - n_{PRI}\right) \log\left(1 - q\right) + \log p\left(q\right) \right] \tag{12}$$

Where

$$\log p\left(q\right) = \log\left(\frac{1}{B\left(\alpha, \beta\right)} q^{\alpha - 1} \left(1 - q\right)^{\beta - 1}\right) \tag{13}$$

The log of $p\left(q\right)$

We have that

$$\log p\left(q\right) = \left(\alpha - 1\right)\log q + \left(\beta - 1\right)\log\left(1 - q\right) - \log B\left(\alpha, \beta\right) \tag{14}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{\left(N - n_{PRI}\right)}{\left(1 - q\right)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

Cinvestav

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1 - q) - \log B(\alpha, \beta) \tag{14}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1 - q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

The log of $p(q)$

We have that

$$\log p(q) = (\alpha - 1)\log q + (\beta - 1)\log(1-q) - \log B(\alpha, \beta) \tag{14}$$

Now taking the derivative with respect to p, we get

$$\frac{n_{PRI}}{q} - \frac{(N - n_{PRI})}{(1-q)} - \frac{\beta - 1}{1 - q} + \frac{\alpha - 1}{q} = 0 \tag{15}$$

Thus

$$\widehat{q}_{MAP} = \frac{n_{PRI} + \alpha - 1}{N + \alpha + \beta - 2} \tag{16}$$

Now

With $N = 20$ with $n_{PRI} = 12$ and $\alpha = \beta = 5$

$$\widehat{q}_{MAP} = 0.571$$

Outline

Cinvestav

Properties

First

MAP estimation "pulls" the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

First

MAP estimation “pulls” the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

First

MAP estimation "pulls" the estimate toward the prior.

Second

The more focused our prior belief, the larger the pull toward the prior.

Example

If $\alpha = \beta =$equal to large value

- It will make the MAP estimate to move closer to the prior.

Properties

Third

In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Cinvestav

Properties

Third

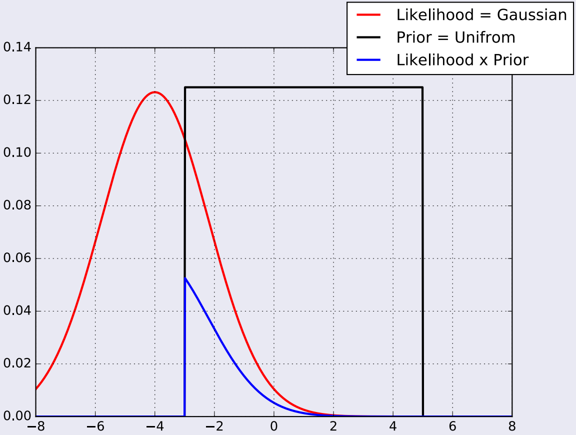
In the expression we derived for $\widehat{q}_{MAP}$, the parameters α and β play a "smoothing" role vis-a-vis the measurement n_{PRI}.

Fourth

Since we referred to q as the parameter to be estimated, we can refer to α and β as the hyper-parameters in the estimation calculations.

Basically the MAP

It is using the power of Likelihood × Prior to obtain more information from the data



Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the **likelihood** $\times$ **the prior**?

For example when we have something like $|\theta_i|$.

We can try the following

EM + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the **likelihood** $\times$ **the prior**?

For example when we have something like $|\theta_i|$.

We can try the following

EM + MAP to be able to estimate the sought parameters.

Beyond simple derivation

In the previous technique

We took an logarithm of the **likelihood** $\times$ **the prior** to obtain a function that can be derived in order to obtain each of the parameters to be estimated.

What if we cannot derive the **likelihood** $\times$ **the prior**?

For example when we have something like $|\theta_i|$.

We can try the following

EM $+$ MAP to be able to estimate the sought parameters.

Outline

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{17}$$

Thus, we have the following probability

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta) \tag{18}$$

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{17}$$

Thus, we have the following probability

$$p(z|\Theta) = p(x, y|\Theta) = p(y|x, \Theta)\, p(x|\Theta) \tag{18}$$

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{17}$$

Thus, we have the following probability

$$p(\boldsymbol{z}|\Theta) = p(\boldsymbol{x}, \boldsymbol{y}|\Theta) = p(\boldsymbol{y}|\boldsymbol{x}, \Theta)\, p(\boldsymbol{x}|\Theta) \tag{18}$$

Cinvestav

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X}, \mathcal{Y}) = \text{Complete Data} \tag{17}$$

Thus, we have the following probability

$$p(\boldsymbol{z}|\Theta) = p(\boldsymbol{x}, \boldsymbol{y}|\Theta) = p(\boldsymbol{y}|\boldsymbol{x}, \Theta)\, p(\boldsymbol{x}|\Theta) \tag{18}$$

Incomplete Data

We assume the following

Two parts of data

1. $\mathcal{X}$ = observed data or **incomplete** data
2. $\mathcal{Y}$ = unobserved data

Thus

$$\mathcal{Z} = (\mathcal{X},\mathcal{Y}) = \text{Complete Data} \tag{17}$$

Thus, we have the following probability

$$p(\boldsymbol{z}|\Theta) = p(\boldsymbol{x},\boldsymbol{y}|\Theta) = p(\boldsymbol{y}|\boldsymbol{x},\Theta)\,p(\boldsymbol{x}|\Theta) \tag{18}$$

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}\left(\Theta|\mathcal{Z}\right)=\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right)=p\left(\mathcal{X},\mathcal{Y}|\Theta\right) \tag{19}$$

Note: The complete data likelihood.

Thus, we have

$$\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right)=p\left(\mathcal{X},\mathcal{Y}|\Theta\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta\right)p\left(\mathcal{X}|\Theta\right) \tag{20}$$

- $p(\mathcal{Y}|\mathcal{X},\Theta)$ is the likelihood of the no observed data under the observed data!!!

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}\left(\Theta|\mathcal{Z}\right) = \mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right) = p\left(\mathcal{X},\mathcal{Y}|\Theta\right) \tag{19}$$

Note: The complete data likelihood.

Thus, we have

$$\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right) = p\left(\mathcal{X},\mathcal{Y}|\Theta\right) = p\left(\mathcal{Y}|\mathcal{X},\Theta\right)p\left(\mathcal{X}|\Theta\right) \tag{20}$$

Did you notice?

- $p\left(\mathcal{X}|\Theta\right)$ is the likelihood of the observed data.
- $p\left(\mathcal{Y}|\mathcal{X},\Theta\right)$ is the likelihood of the no-observed data under the observed data!!!

New Likelihood Function

The New Likelihood Function

$$\mathcal{L}(\Theta|\mathcal{Z}) = \mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\Theta) \tag{19}$$

Note: The complete data likelihood.

Thus, we have

$$\mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = p(\mathcal{X},\mathcal{Y}|\Theta) = p(\mathcal{Y}|\mathcal{X},\Theta)\,p(\mathcal{X}|\Theta) \tag{20}$$

Did you notice?

- $p(\mathcal{X}|\Theta)$ is the likelihood of the observed data.
- $p(\mathcal{Y}|\mathcal{X},\Theta)$ is the likelihood of the no-observed data under the observed data!!!

Rewriting

This can be rewritten as

$$\mathcal{L}(\Theta|\mathcal{X},\mathcal{Y}) = h_{\mathcal{X},\Theta}(\mathcal{Y}) \tag{21}$$

This basically signify that $\mathcal{X}, \Theta$ are constant and the only random part is $\mathcal{Y}$.

In addition

$$\mathcal{L}(\Theta|\mathcal{X}) \tag{22}$$

It is known as the incomplete-data likelihood function.

Cinvestav

Rewriting

This can be rewritten as

$$\mathcal{L}\left(\Theta|\mathcal{X},\mathcal{Y}\right) = h_{\mathcal{X},\Theta}\left(\mathcal{Y}\right) \tag{21}$$

This basically signify that $\mathcal{X},\Theta$ are constant and the only random part is $\mathcal{Y}$.

In addition

$$\mathcal{L}\left(\Theta|\mathcal{X}\right) \tag{22}$$

It is known as the incomplete-data likelihood function.

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) =& p(\mathcal{X}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta) \\
=& \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \\
=& \sum \left(\prod_{i=1}^{N} p(x_i|\Theta)\right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Cinvestav

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) &= p(\mathcal{X}|\Theta) \\
&= \sum_{\mathcal{Y}} p(\mathcal{X},\mathcal{Y}|\Theta) \\
&= \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)\, p(\mathcal{X}|\Theta) \\
&= \sum \left(\prod_{i=1}^{N} p(x_i|\Theta)\right)_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X},\Theta)
\end{aligned}
$$

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}(\Theta|\mathcal{X}) &= p(\mathcal{X}|\Theta) \\
&= \sum_{\mathcal{Y}} p(\mathcal{X}, \mathcal{Y}|\Theta) \\
&= \sum_{\mathcal{Y}} p(\mathcal{Y}|\mathcal{X}, \Theta)\, p(\mathcal{X}|\Theta) \\
&= \sum_{\mathcal{Y}} \left(\prod_{i=1}^{N} p(x_i|\Theta)\right) p(\mathcal{Y}|\mathcal{X}, \Theta)
\end{aligned}
$$

Thus

We can connect both incomplete-complete data equations by doing the following

$$
\begin{aligned}
\mathcal{L}\left(\Theta|\mathcal{X}\right) =& p\left(\mathcal{X}|\Theta\right)\\
=& \sum_{\mathcal{Y}} p\left(\mathcal{X},\mathcal{Y}|\Theta\right)\\
=& \sum_{\mathcal{Y}} p\left(\mathcal{Y}|\mathcal{X},\Theta\right) p\left(\mathcal{X}|\Theta\right)\\
=& \sum \left(\prod_{i=1}^{N} p\left(x_i|\Theta\right)\right)_{\mathcal{Y}} p\left(\mathcal{Y}|\mathcal{X},\Theta\right)
\end{aligned}
$$

Remarks

Problems

Normally, it is almost impossible to obtain a closed analytical solution for the previous equation.

However

We can use the expected value of $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$, which allows us to find an iterative procedure to approximate the solution.

Remarks

Problems

Normally, it is almost impossible to obtain a closed analytical solution for the previous equation.

However

We can use the expected value of $\log p(\mathcal{X}, \mathcal{Y}|\Theta)$, which allows us to find an iterative procedure to approximate the solution.

The function we would like to have

The Q function

We want an estimation of the complete-data log-likelihood

$$\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right) \tag{23}$$

Based in the info provided by $\mathcal{X}, \Theta_{n-1}$ where Θ_{n-1} is a previously estimated set of parameters at step n.

Think about the following, if we want to remove $\mathcal{Y}$

$$\int \left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)\right] p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right) d\mathcal{Y} \tag{24}$$

Remark: We integrate out $\mathcal{Y}$ - Actually, this is the expected value of $\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)$.

Cinvestav

The function we would like to have

The Q function

We want an estimation of the complete-data log-likelihood

$$\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right) \tag{23}$$

Based in the info provided by $\mathcal{X}, \Theta_{n-1}$ where Θ_{n-1} is a previously estimated set of parameters at step n.

Think about the following, if we want to remove $\mathcal{Y}$

$$\int \left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)\right] p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right) d\mathcal{Y} \tag{24}$$

Remark: We integrate out $\mathcal{Y}$ - Actually, this is the expected value of $\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)$.

Outline

Cinvestav

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)|\mathcal{X}, \Theta_{n-1}\right] \tag{25}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Cinvestav

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y} | \Theta\right) | \mathcal{X}, \Theta_{n-1}\right] \tag{25}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y} | \Theta\right) | \mathcal{X}, \Theta_{n-1}\right] \tag{25}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Use the Expected Value

Then, we want an iterative method to guess Θ from Θ_{n-1}

$$Q\left(\Theta, \Theta_{n-1}\right) = E\left[\log p\left(\mathcal{X}, \mathcal{Y}|\Theta\right)|\mathcal{X}, \Theta_{n-1}\right] \tag{25}$$

Take in account that

1. $\mathcal{X}, \Theta_{n-1}$ are taken as constants.
2. Θ is a normal variable that we wish to adjust.
3. $\mathcal{Y}$ is a random variable governed by distribution $p\left(\mathcal{Y}|\mathcal{X}, \Theta_{n-1}\right)$=marginal distribution of missing data.

Cinvestav

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\boldsymbol{\mathcal{Y}}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \tag{26}$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\boldsymbol{\mathcal{Y}}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \tag{26}$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\boldsymbol{\mathcal{Y}}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \tag{26}$$

which is not dependent on Θ.

Another Interpretation

Given the previous information

$$E\left[\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)|\mathcal{X},\Theta_{n-1}\right]=\int_{\boldsymbol{\mathcal{Y}}\in\mathbb{Y}}\log p\left(\mathcal{X},\mathcal{Y}|\Theta\right)p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)d\mathcal{Y}$$

Something Notable

1. In the best of cases, this marginal distribution is a simple analytical expression of the assumed parameter Θ_{n-1}.
2. In the worst of cases, this density might be very hard to obtain.

Actually, we use

$$p\left(\mathcal{Y},\mathcal{X}|\Theta_{n-1}\right)=p\left(\mathcal{Y}|\mathcal{X},\Theta_{n-1}\right)p\left(\mathcal{X}|\Theta_{n-1}\right) \quad (26)$$

which is not dependent on Θ.

Outline

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, Y)$ a function
  - θ a constant
  - $Y \sim p_Y(y)$, a random variable with distribution $p_Y(y)$.

Thus, if Y is a discrete random variable

$$q(\theta) = E_Y[h(\theta, Y)] = \sum_y h(\theta, y)\, p_Y(y) \tag{27}$$

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \tag{27}$$

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}[h(\theta, \boldsymbol{Y})] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \tag{27}$$

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\boldsymbol{Y}}(y)$, a random variable with distribution $p_{\boldsymbol{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\boldsymbol{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\boldsymbol{Y}}(y) \tag{27}$$

Cinvestav

Back to the Q function

The intuition

We have the following analogy:

- Consider $h(\theta, \boldsymbol{Y})$ a function
  - θ a constant
  - $\boldsymbol{Y} \sim p_{\mathbf{Y}}(y)$, a random variable with distribution $p_{\mathbf{Y}}(y)$.

Thus, if $\boldsymbol{Y}$ is a discrete random variable

$$q(\theta) = E_{\mathbf{Y}}\left[h(\theta, \boldsymbol{Y})\right] = \sum_{y} h(\theta, y)\, p_{\mathbf{Y}}(y) \tag{27}$$

Cinvestav

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \tag{28}$$

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \quad (28)$$

Cinvestav

Why E-step!!!

From here the name

This is basically the E-step

The second step

It tries to maximize the Q function

$$\Theta_n = \text{argmax}_{\Theta} Q\left(\Theta, \Theta_{n-1}\right) \tag{28}$$

The EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{29}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{30}$$

Cinvestav

The EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{29}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{30}$$

The EM-Algorithm

The likelihood function we are going to use

Let $\mathcal{X}$ be a random vector which results from a parametrized family:

$$\mathcal{L}(\Theta) = \ln \mathcal{P}(\mathcal{X}|\Theta) \tag{29}$$

Note: $\ln(x)$ is a strictly increasing function.

We wish to compute Θ

Based on an estimate Θ_n (After the n^{th}) such that $\mathcal{L}(\Theta) > \mathcal{L}(\Theta_n)$

Or the maximization of the difference

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln \mathcal{P}(\mathcal{X}|\Theta) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \tag{30}$$

Cinvestav

Outline

Cinvestav

Introducing the Hidden Features

Given that the hidden random vector $\mathcal{Y}$ exits with y values

$$\mathcal{P}\left(\mathcal{X}|\Theta\right) = \sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \tag{31}$$

Thus, using our first constraint $\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right)$

$$\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) = \ln\left(\sum_{y} \mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right) - \ln\mathcal{P}\left(\mathcal{X}|\Theta_n\right) \tag{32}$$

Introducing the Hidden Features

Given that the hidden random vector $\mathcal{Y}$ exits with y values

$$\mathcal{P}\left(\mathcal{X}|\Theta\right)=\sum_{y}\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \tag{31}$$

Thus, using our first constraint $\mathcal{L}\left(\Theta\right)-\mathcal{L}\left(\Theta_n\right)$

$$\mathcal{L}\left(\Theta\right)-\mathcal{L}\left(\Theta_n\right)=\ln\left(\sum_{y}\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right)-\ln\mathcal{P}\left(\mathcal{X}|\Theta_n\right) \tag{32}$$

Cinvestav

Here, we introduce some concepts of convexity

For Convexity

Theorem (Jensen's inequality)

Let f be a convex function defined on an interval I. If $x_1, x_2, ..., x_n \in I$ and $\lambda_1, \lambda_2, ..., \lambda_n \geq 0$ with $\sum_{i=1}^{n} \lambda_i = 1$, then

$$f\left(\sum_{i=1}^{n} \lambda_i x_i\right) \leq \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \tag{33}$$

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Proof:

For $n = 1$

We have the trivial case

For $n = 2$

The convexity definition.

Now the inductive hypothesis

We assume that the theorem is true for some n.

Now, we have

The following linear combination for λ_i

$$f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) = f\left(\lambda_{n+1}x_{n+1} + \sum_{i=1}^{n}\lambda_i x_i\right)$$

$$= f\left(\lambda_{n+1}x_{n+1} + \frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)$$

$$\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{(1-\lambda_{n+1})}\sum_{i=1}^{n}\lambda_i x_i\right)$$

Now, we have

The following linear combination for λ_i

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right) \\
&= f\left(\lambda_{n+1} x_{n+1} + \frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right)
\end{aligned}
$$

Now, we have

The following linear combination for λ_i

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &= f\left(\lambda_{n+1} x_{n+1} + \sum_{i=1}^{n} \lambda_i x_i\right) \\
&= f\left(\lambda_{n+1} x_{n+1} + \frac{(1-\lambda_{n+1})}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i x_i\right)
\end{aligned}
$$

Cinvestav

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1-\lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Did you notice?

Something Notable

$$\sum_{i=1}^{n+1} \lambda_i = 1$$

Thus

$$\sum_{i=1}^{n} \lambda_i = 1 - \lambda_{n+1}$$

Finally

$$\frac{1}{(1 - \lambda_{n+1})} \sum_{i=1}^{n} \lambda_i = 1$$

Cinvestav

Now

We have that

$$f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) \leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i x_i\right)$$

$$\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) \frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i f\left(x_i\right)$$

$$\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n}\lambda_i f\left(x_i\right) \text{ Q.E.D.}$$

Now

We have that

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1} \lambda_i x_i\right) &\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) \frac{1}{\left(1-\lambda_{n+1}\right)} \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n} \lambda_i f\left(x_i\right) \text{ Q.E.D.}
\end{aligned}
$$

Now

We have that

$$
\begin{aligned}
f\left(\sum_{i=1}^{n+1}\lambda_i x_i\right) &\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right) f\left(\frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i x_i\right)\\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \left(1-\lambda_{n+1}\right)\frac{1}{\left(1-\lambda_{n+1}\right)}\sum_{i=1}^{n}\lambda_i f\left(x_i\right)\\
&\leq \lambda_{n+1} f\left(x_{n+1}\right) + \sum_{i=1}^{n}\lambda_i f\left(x_i\right) \text{ Q.E.D.}
\end{aligned}
$$

Cinvestav

Thus, for concave functions

It is possible to shown that

Given $\ln(x)$ a concave function:

$$\ln\left[\sum_{i=1}^{n} \lambda_i x_i\right] \geq \sum_{i=1}^{n} \lambda_i \ln(x_i)$$

If we take in consideration

Assume that the $\lambda_i = \mathcal{P}(y|\mathcal{X}, \Theta_n)$. We know that

1. $\mathcal{P}(y|\mathcal{X}, \Theta_n) \geq 0$
2. $\sum_y \mathcal{P}(y|\mathcal{X}, \Theta_n) = 1$

Thus, for concave functions

It is possible to shown that

Given $\ln(x)$ a concave function:

$$\ln\left[\sum_{i=1}^{n}\lambda_i x_i\right] \geq \sum_{i=1}^{n}\lambda_i \ln(x_i)$$

If we take in consideration

Assume that the $\lambda_i = \mathcal{P}(y|\mathcal{X},\Theta_n)$. We know that

1. $\mathcal{P}(y|\mathcal{X},\Theta_n) \geq 0$
2. $\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) = 1$

We have

First

$$\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) = \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n)$$

$$\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln \mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) &= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln\mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln\mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln\mathcal{P}(\mathcal{X}|\Theta_n) \\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots
\end{aligned}
$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta) - \mathcal{L}(\Theta_n) &= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \ln \mathcal{P}(\mathcal{X}|\Theta_n) \\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right) - \cdots
\end{aligned}
$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n) \ln \mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

We have

First

$$
\begin{aligned}
\mathcal{L}(\Theta)-\mathcal{L}(\Theta_n) &= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\right)-\ln\mathcal{P}(\mathcal{X}|\Theta_n)\\
&= \ln\left(\sum_y \mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)\,\frac{\mathcal{P}(y|\mathcal{X},\Theta_n)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right)-\ln\mathcal{P}(\mathcal{X}|\Theta_n)\\
&= \ln\left(\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\,\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right)-\ln\mathcal{P}(\mathcal{X}|\Theta_n)\\
&\geq \sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)}\right)-\dots
\end{aligned}
$$

$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\mathcal{P}(\mathcal{X}|\Theta_n)$ Why this?

Next

Because

$$\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) = 1$$

Then

$$\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) \geq \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Next

Because

$$\sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) = 1$$

Then

$$\begin{aligned}\mathcal{L}\left(\Theta\right) - \mathcal{L}\left(\Theta_n\right) \geq & \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\ = & \Delta\left(\Theta|\Theta_n\right)\end{aligned}$$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \quad (34)$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \quad (35)$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{34}$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{35}$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Then, we have

Then, we have proved that

$$\mathcal{L}(\Theta) \geq \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{34}$$

Then, we define a new function

$$l(\Theta|\Theta_n) = \mathcal{L}(\Theta_n) + \Delta(\Theta|\Theta_n) \tag{35}$$

Thus $l(\Theta|\Theta_n)$

It is bounded from above by $\mathcal{L}(\Theta)$ i.e $l(\Theta|\Theta_n) \leq \mathcal{L}(\Theta)$

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right)+\Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right)+\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)+\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta=\Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right)\\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Cinvestav

Now, we can do the following

We evaluate in Θ_n

$$
\begin{aligned}
l\left(\Theta_n|\Theta_n\right) =& \mathcal{L}\left(\Theta_n\right) + \Delta\left(\Theta_n|\Theta_n\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta_n\right)\mathcal{P}\left(y|\Theta_n\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}{\mathcal{P}\left(\mathcal{X},y|\Theta_n\right)}\right) \\
=& \mathcal{L}\left(\Theta_n\right)
\end{aligned}
$$

This means that

For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

Cinvestav

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Therefore

The function $l\left(\Theta|\Theta_n\right)$ has the following properties

1. It is bounded from above by $\mathcal{L}\left(\Theta\right)$ i.e $l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$.
2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal.
3. The function $l\left(\Theta|\Theta_n\right)$ is concave... How?

Outline

Cinvestav

First

We have the value $\mathcal{L}\left(\Theta_n\right)$

We know that $\mathcal{L}\left(\Theta_n\right)$ is constant i.e. an offset value

What about $\Delta\left(\Theta|\Theta_n\right)$

$$\sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

We have that the ln is a concave function

$$\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Cinvestav

First

We have the value $\mathcal{L}(\Theta_n)$

We know that $\mathcal{L}(\Theta_n)$ is constant i.e. an offset value

What about $\Delta(\Theta|\Theta_n)$

$$\sum_y \mathcal{P}(y|\mathcal{X},\Theta_n)\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

We have that the ln is a concave function

$$\ln\left(\frac{\mathcal{P}(\mathcal{X}|y,\Theta)\,\mathcal{P}(y|\Theta)}{\mathcal{P}(y|\mathcal{X},\Theta_n)\,\mathcal{P}(\mathcal{X}|\Theta_n)}\right)$$

First

We have the value $\mathcal{L}\left(\Theta_n\right)$

We know that $\mathcal{L}\left(\Theta_n\right)$ is constant i.e. an offset value

What about $\Delta\left(\Theta|\Theta_n\right)$

$$\sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

We have that the $\ln$ is a concave function

$$\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Cinvestav

Therefore

Each element is concave

$$\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore, the sum of concave functions is a concave function

$$\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore

Each element is concave

$$\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Therefore, the sum of concave functions is a concave function

$$\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)$$

Cinvestav

Outline

Cinvestav

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l\left(\Theta|\Theta_n\right)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l(\Theta|\Theta_n)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph

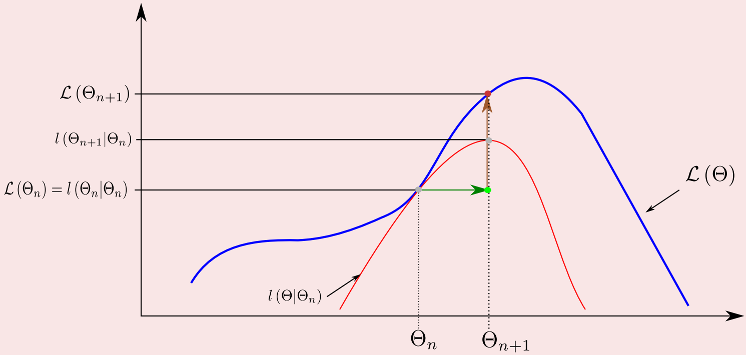
Cinvestav

Given the Concave Function

Thus, we have that

1. We can select Θ_n such that $l\left(\Theta|\Theta_n\right)$ is maximized.
2. Thus, given a Θ_n, we can generate Θ_{n+1}.

The process can be seen in the following graph



Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Given

The Previous Constraints

1. $l\left(\Theta|\Theta_n\right)$ is bounded from above by $\mathcal{L}\left(\Theta\right)$

$$l\left(\Theta|\Theta_n\right) \leq \mathcal{L}\left(\Theta\right)$$

2. For $\Theta = \Theta_n$, functions $\mathcal{L}\left(\Theta\right)$ and $l\left(\Theta|\Theta_n\right)$ are equal

$$\mathcal{L}\left(\Theta_n\right) = l\left(\Theta|\Theta_n\right)$$

3. The function $l\left(\Theta|\Theta_n\right)$ is concave

Outline

Cinvestav

From

The following

$$
\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)} \right) \right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
\approx& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} = & \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
\approx & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\} \\
= & \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)} \right) \right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
\approx& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right)\right\} \\
=& \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)}\right)\right\}
\end{aligned}
$$

The terms with Θ_n are constants.

$$
\begin{aligned}
\approx& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)\right)\right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_y \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}
\end{aligned}
$$

From

The following

$$
\begin{aligned}
\Theta_{n+1} =& \text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \mathcal{L}\left(\Theta_n\right) + \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right)}{\mathcal{P}\left(y|\mathcal{X},\Theta_n\right)\mathcal{P}\left(\mathcal{X}|\Theta_n\right)} \right) \right\} \\
& \text{The terms with } \Theta_n \text{ are constants.} \\
\approx& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\mathcal{P}\left(\mathcal{X}|y,\Theta\right)\mathcal{P}\left(y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X},y|\Theta\right)}{\mathcal{P}\left(y|\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\frac{\mathcal{P}(\mathcal{X},y,\Theta)}{\mathcal{P}(\Theta)}}{\frac{\mathcal{P}(y,\Theta)}{\mathcal{P}(\Theta)}} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}
\end{aligned}
$$

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta}\left\{l\left(\Theta|\Theta_n\right)\right\} \approx \text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}$

Thus

Then

$$\theta_{n+1} = \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X}, y, \Theta\right)}{\mathcal{P}\left(y, \Theta\right)} \frac{\mathcal{P}\left(y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}$$

$$= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X}, y, \Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\}$$

$$= \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X}, \Theta_n\right) \ln \left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right\}$$

$$= \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X}, \Theta_n} \left[\ln \left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right] \right\}$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X}, \Theta_n} \left[\ln \left(\mathcal{P}\left(\mathcal{X}, y|\Theta\right) \right) \right] \right\}$

Cinvestav

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln \left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n} \left[\ln \left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n} \left[\ln \left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right] \right\}$

Thus

Then

$$
\begin{aligned}
\theta_{n+1} =& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)} \frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)} \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ \sum_{y} \mathcal{P}\left(y|\mathcal{X},\Theta_n\right) \ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right) \right) \right\} \\
=& \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right] \right\}
\end{aligned}
$$

Then $\text{argmax}_{\Theta} \left\{ l\left(\Theta|\Theta_n\right) \right\} \approx \text{argmax}_{\Theta} \left\{ E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right] \right\}$

Thus

Then

$$\begin{aligned}
\theta_{n+1} =&\text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(y,\Theta\right)}\frac{\mathcal{P}\left(y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}\\
=&\text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\frac{\mathcal{P}\left(\mathcal{X},y,\Theta\right)}{\mathcal{P}\left(\Theta\right)}\right)\right\}\\
=&\text{argmax}_{\Theta}\left\{\sum_{y}\mathcal{P}\left(y|\mathcal{X},\Theta_{n}\right)\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right\}\\
=&\text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_{n}}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}
\end{aligned}$$

Then $\text{argmax}_{\Theta}\left\{l\left(\Theta|\Theta_{n}\right)\right\}\approx\text{argmax}_{\Theta}\left\{E_{y|\mathcal{X},\Theta_{n}}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]\right\}$

Outline

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

Cinvestav

The EM-Algorithm

Steps of EM

1. Expectation under hidden variables.
2. Maximization of the resulting formula.

E-Step

Determine the conditional expectation, $E_{y|\mathcal{X},\Theta_n}\left[\ln\left(\mathcal{P}\left(\mathcal{X},y|\Theta\right)\right)\right]$.

M-Step

Maximize this expression with respect to Θ.

Outline

Cinvestav

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that Θ_{n+1} is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that $\Theta_{n+1}$is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Cinvestav

Notes and Convergence of EM

Gains between $\mathcal{L}(\Theta)$ and $l(\Theta|\Theta_n)$

Using the hidden variables it is possible to simplify the optimization of $\mathcal{L}(\Theta)$ through $l(\Theta|\Theta_n)$.

Convergence

- Remember that $\Theta_{n+1}$is the estimate for Θ which maximizes the difference $\Delta(\Theta|\Theta_n)$.

Cinvestav

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right) = 0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right) \geq \Delta\left(\Theta_n|\Theta_n\right) = 0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Cinvestav

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right) = 0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right) \geq \Delta\left(\Theta_n|\Theta_n\right) = 0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Cinvestav

Therefore

Then, we have

Given the initial estimate of Θ by Θ_n

$$\Delta\left(\Theta_n|\Theta_n\right) = 0$$

Now

If we choose Θ_{n+1} to maximize the $\Delta\left(\Theta|\Theta_n\right)$, then

$$\Delta\left(\Theta_{n+1}|\Theta_n\right) \geq \Delta\left(\Theta_n|\Theta_n\right) = 0$$

We have that

The Likelihood $\mathcal{L}\left(\Theta\right)$ is not a decreasing function with respect to Θ.

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(x\right)=x$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right]=\Theta^*$$

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(\boldsymbol{x}\right) = \boldsymbol{x}$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right] = \Theta^*$$

Cinvestav

Notes and Convergence of EM

Properties

When the algorithm reaches a fixed point for some Θ_n, the value maximizes $l\left(\Theta|\Theta_n\right)$.

Definition

A fixed point of a function is an element on domain that is mapped to itself by the function:

$$f\left(\boldsymbol{x}\right) = \boldsymbol{x}$$

Basically the EM algorithm does the following

$$EM\left[\Theta^*\right] = \Theta^*$$

Cinvestav

At this moment

We have that

The algorithm reaches a fixed point for some Θ_n, the value Θ^* maximizes $l\left(\Theta|\Theta_n\right)$.

Then, when the algorithm

- It reaches a fixed point for some Θ_n the value maximizes $l\left(\Theta|\Theta_n\right)$.
  - Basically $\Theta_{n+1} = \Theta_n$.

Cinvestav

At this moment

We have that

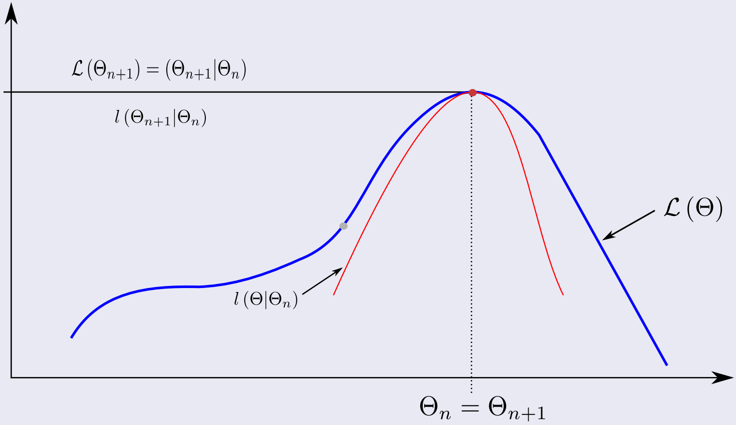
The algorithm reaches a fixed point for some Θ_n, the value Θ^* maximizes $l\left(\Theta|\Theta_n\right)$.

Then, when the algorithm

- It reaches a fixed point for some Θ_n the value maximizes $l\left(\Theta|\Theta_n\right)$.
  - Basically $\Theta_{n+1} = \Theta_n$.

Therefore

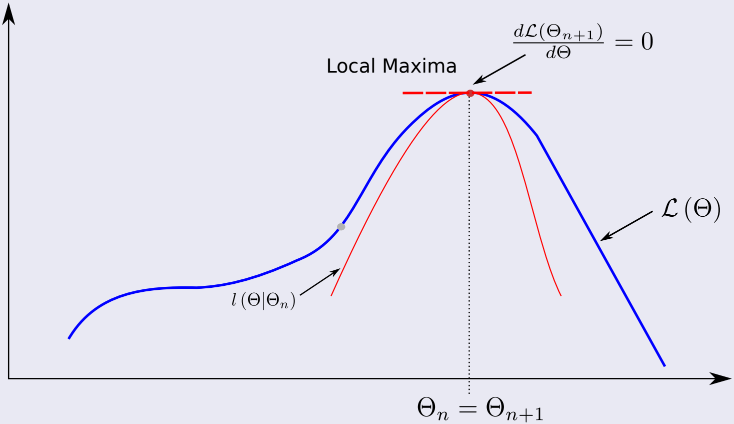
We have



Then

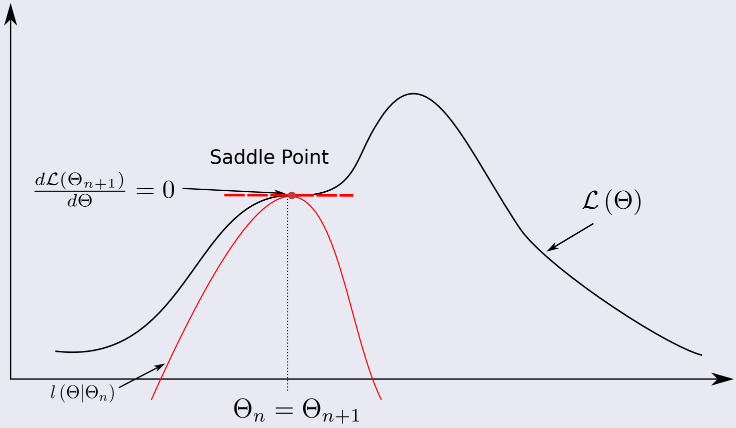
If $\mathcal{L}$ and l are differentiable at Θ_n

- Since $\mathcal{L}$ and l are equal at Θ_n
  - Then, Θ_n is a stationary point of $\mathcal{L}$ i.e. the derivative of $\mathcal{L}$ vanishes at that point.



However

You could finish with the following case, no local maxima



For more on the subject

Please take a look to

Geoffrey McLachlan and Thriyambakam Krishnan, *"The EM Algorithm and Extensions,"* John Wiley & Sons, New York, 1996.

Outline

Cinvestav

Example

This application comes from

"Adaptive Sparseness for Supervised Learning" by Mário A.T. Figueiredo

In

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 25, NO. 9, SEPTEMBER 2003

Example

This application comes from

"Adaptive Sparseness for Supervised Learning" by Mário A.T. Figueiredo

In

IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, VOL. 25, NO. 9, SEPTEMBER 2003

Outline

Cinvestav

Linear Regression with Gaussian Prior

We consider regression functions that are linear with respect to the parameter vector β

$$f(\boldsymbol{x}, \boldsymbol{w}) = \sum_{i=1}^{k} w_i h(x) = \boldsymbol{w}^T \boldsymbol{h}(\boldsymbol{x})$$

Where

$\boldsymbol{h}(\boldsymbol{x}) = [h_1(\boldsymbol{x}), ..., h_k(\boldsymbol{x})]^T$ is a vector of k fixed function of the input, often called features.

Linear Regression with Gaussian Prior

We consider regression functions that are linear with respect to the parameter vector β

$$f\left(\boldsymbol{x},\boldsymbol{w}\right)=\sum_{i=1}^{k} w_i h\left(x\right)=\boldsymbol{w}^T\boldsymbol{h}\left(\boldsymbol{x}\right)$$

Where

$\boldsymbol{h}\left(\boldsymbol{x}\right)=\left[h_1\left(\boldsymbol{x}\right),...,h_k\left(\boldsymbol{x}\right)\right]^T$ is a vector of k fixed function of the input, often called features.

Actually, it can be...

Linear Regression

Linear regression, in which $\boldsymbol{h}(\boldsymbol{x}) = [1, x_1, ..., x_d]^T$ i; in this case, $k = d + 1$.

Non-Linear Regression

Here, you have a fixed basis function where $\boldsymbol{h}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), ..., \phi_1(\boldsymbol{x})]^T$ with $\phi_1(\boldsymbol{x}) = 1$.

Kernel Regression

Here $\boldsymbol{h}(\boldsymbol{x}) = [1, K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), ..., K(\boldsymbol{x}, \boldsymbol{x}_n)]^T$ where $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is some kernel function.

Cinvestav

Actually, it can be...

Linear Regression

Linear regression, in which $\boldsymbol{h}(\boldsymbol{x}) = [1, x_1, ..., x_d]^T$ i; in this case, $k = d + 1$.

Non-Linear Regression

Here, you have a fixed basis function where $\boldsymbol{h}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), ..., \phi_1(\boldsymbol{x})]^T$ with $\phi_1(\boldsymbol{x}) = 1$.

Kernel Regression

Here $\boldsymbol{h}(\boldsymbol{x}) = [1, K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), ..., K(\boldsymbol{x}, \boldsymbol{x}_n)]^T$ where $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is some kernel function.

Actually, it can be...

Linear Regression

Linear regression, in which $\boldsymbol{h}(\boldsymbol{x}) = [1, x_1, ..., x_d]^T$ i; in this case, $k = d + 1$.

Non-Linear Regression

Here, you have a fixed basis function where
$\boldsymbol{h}(\boldsymbol{x}) = [\phi_1(\boldsymbol{x}), \phi_2(\boldsymbol{x}), ..., \phi_1(\boldsymbol{x})]^T$ with $\phi_1(\boldsymbol{x}) = 1$.

Kernel Regression

Here $\boldsymbol{h}(\boldsymbol{x}) = [1, K(\boldsymbol{x}, \boldsymbol{x}_1), K(\boldsymbol{x}, \boldsymbol{x}_2), ..., K(\boldsymbol{x}, \boldsymbol{x}_n)]^T$ where $K(\boldsymbol{x}, \boldsymbol{x}_i)$ is some kernel function.

Outline

Cinvestav

Gaussian Noise

We assume that the training set is contaminated by additive white Gaussian Noise

$$y_i = f(\boldsymbol{x}_i, \boldsymbol{w}) + \omega_i = \boldsymbol{w}^T \boldsymbol{x}_i + \omega_i \tag{36}$$

for $i = 1, ..., N$ where $[\omega_1, ..., \omega_N]$ is a set of independent zero-mean Gaussian samples with variance σ^2

With $f(\boldsymbol{x}_i, \boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{x}_i$

Thus, for $\boldsymbol{y} = [y_1, ..., y_N]^T$, we have the following likelihood

$$p(\omega_1, \omega_2, ..., \omega_N) = \prod_{i=1}^{N} p\left(\omega_i | 0, \sigma^2\right)$$

Gaussian Noise

We assume that the training set is contaminated by additive white Gaussian Noise

$$y_i = f\left(\boldsymbol{x}_i, \boldsymbol{w}\right) + \omega_i = \boldsymbol{w}^T \boldsymbol{x}_i + \omega_i \quad (36)$$

for $i = 1, ..., N$ where $[\omega_1, ..., \omega_N]$ is a set of independent zero-mean Gaussian samples with variance σ^2

With $f\left(\boldsymbol{x}_i, \boldsymbol{w}\right) = \boldsymbol{w}^T \boldsymbol{x}_i$

Thus, for $\boldsymbol{y} = [y_1, ..., y_N]^T$, we have the following likelihood

$$p\left(\omega_1, \omega_2, ..., \omega_N\right) = \prod_{i=1}^{N} p\left(\omega_i | 0, \sigma^2\right)$$

Something Interesting

We have that

$$\prod_{i=1}^{N} p\left(\omega_i|0,\sigma^2\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\omega_i^2}{2\sigma^2}\right\}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\left(y_i - \boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Therefore

$$\frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\left(y_i - \boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\exp\left\{-\sum_{i=1}^{N}\frac{\left(y_i - \boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Cinvestav

Something Interesting

We have that

$$\prod_{i=1}^{N} p\left(\omega_i | 0, \sigma^2\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\,\sigma^N} \prod_{i=1}^{N} \exp\left\{-\frac{\omega_i^2}{2\sigma^2}\right\}$$

$$= \frac{1}{(2\pi)^{\frac{N}{2}}\,\sigma^N} \prod_{i=1}^{N} \exp\left\{-\frac{\left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Therefore

$$\frac{1}{(2\pi)^{\frac{N}{2}}\,\sigma^N} \prod_{i=1}^{N} \exp\left\{-\frac{\left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{\frac{N}{2}}\,\sigma^N} \exp\left\{-\sum_{i=1}^{N} \frac{\left(y_i - \boldsymbol{w}^T \boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Cinvestav

Something Interesting

We have that

$$\prod_{i=1}^{N} p\left(\omega_i|0,\sigma^2\right) = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\omega_i^2}{2\sigma^2}\right\}$$
$$= \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\left(y_i-\boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Therefore

$$\frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\prod_{i=1}^{N}\exp\left\{-\frac{\left(y_i-\boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\} = \frac{1}{(2\pi)^{\frac{N}{2}}\sigma^N}\exp\left\{-\sum_{i=1}^{N}\frac{\left(y_i-\boldsymbol{w}^T\boldsymbol{x}_i\right)^2}{2\sigma^2}\right\}$$

Cinvestav

Then

We can rewrite this in vector form

$$p\left(\boldsymbol{y}|X\boldsymbol{w},\sigma^2 I\right) \approx \exp\left\{-\left(\boldsymbol{y}-X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y}-X\boldsymbol{w}\right)\right\}$$

- With $\sigma' = \sqrt{2}\sigma$

Thus, for $[y_1, \ldots, y_N]$, we have the following likelihood

$$p\left(\boldsymbol{y}|\boldsymbol{w}\right) = \mathcal{N}\left(X\boldsymbol{w}, \sigma'^2 I\right) \quad (37)$$

Then

We can rewrite this in vector form

$$p\left(\boldsymbol{y}|X\boldsymbol{w},\sigma^2 I\right) \approx \exp\left\{-\left(\boldsymbol{y}-X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y}-X\boldsymbol{w}\right)\right\}$$

- With $\sigma' = \sqrt{2}\sigma$

Thus, for $[y_1, ..., y_N]$, we have the following likelihood

$$p\left(\boldsymbol{y}|\boldsymbol{w}\right) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma'^2 I\right) \quad (37)$$

Gaussian Noise

What if we assume a prior zero mean Gaussian for $\boldsymbol{w}$

$$p\left(\boldsymbol{w}|0, A\right) = N\left(0, A\right)$$

The posterior looks like

$$p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx \exp\left\{-\left(\boldsymbol{y} - X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y} - X\boldsymbol{w}\right)\right\} \exp\left\{-\boldsymbol{w}^T A^{-1} \boldsymbol{w}\right\} \quad (38)$$

We have the following

$$\log p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx -\left(\boldsymbol{y} - X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y} - X\boldsymbol{w}\right) - \boldsymbol{w}^T A^{-1} \boldsymbol{w}$$

Gaussian Noise

What if we assume a prior zero mean Gaussian for $\boldsymbol{w}$

$$p\left(\boldsymbol{w}|0, A\right) = N\left(0, A\right)$$

The posterior looks like

$$p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx \exp\left\{-\left(\boldsymbol{y}-X\boldsymbol{w}\right)^{T}\frac{1}{\sigma'^{2}}I\left(\boldsymbol{y}-X\boldsymbol{w}\right)\right\}\exp\left\{-\boldsymbol{w}^{T}A^{-1}\boldsymbol{w}\right\} \quad (38)$$

We have the following

$$\log p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx -\left(\boldsymbol{y}-X\boldsymbol{w}\right)^{T}\frac{1}{\sigma'^{2}}I\left(\boldsymbol{y}-X\boldsymbol{w}\right)-\boldsymbol{w}^{T}A^{-1}\boldsymbol{w}$$

Gaussian Noise

What if we assume a prior zero mean Gaussian for $\boldsymbol{w}$

$$p\left(\boldsymbol{w}|0, A\right) = N\left(0, A\right)$$

The posterior looks like

$$p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx \exp\left\{-\left(\boldsymbol{y} - X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y} - X\boldsymbol{w}\right)\right\} \exp\left\{-\boldsymbol{w}^T A^{-1} \boldsymbol{w}\right\} \quad (38)$$

We have the following

$$\log p\left(\boldsymbol{w}|\boldsymbol{y}\right) \approx -\left(\boldsymbol{y} - X\boldsymbol{w}\right)^T \frac{1}{\sigma'^2} I \left(\boldsymbol{y} - X\boldsymbol{w}\right) - \boldsymbol{w}^T A^{-1} \boldsymbol{w}$$

Therefore

The posterior $p\left(\boldsymbol{w}|\boldsymbol{y}\right)$ is still Gaussian and the mode/maximal estimation is given by

$$\widehat{\boldsymbol{w}} = \left(\sigma^2 A^{-1} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \tag{39}$$

Remark: The Ridge regression.

Outline

Cinvestav

Regression with a Laplacian Prior

Thus, the MAP estimate of $\boldsymbol{w}$ look like

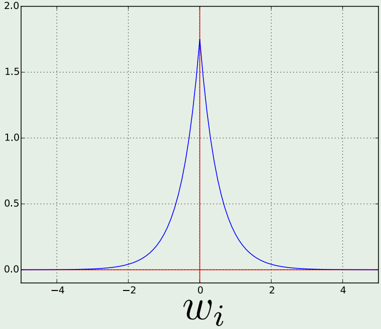
$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + 2\sigma^2 \alpha \left\| \boldsymbol{w} \right\|_1 \right\} \quad (40)$$

Cinvestav

Regression with a Laplacian Prior

In order to favor sparse estimate, we can adopt priors

$$p(\boldsymbol{w}|\alpha) = \prod_{i=1}^{d} \frac{\alpha}{2} \exp\{-\alpha |w_i|\} = \left(\frac{\alpha}{2}\right)^d \exp\{-\alpha \|\boldsymbol{w}\|_1\} \quad (41)$$



Regression with a Laplacian Prior

Thus, the Maximum A Posterior (MAP) estimate of $\boldsymbol{w}$ look like

$$\widehat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + 2\sigma^2 \alpha \left\| \boldsymbol{w} \right\|_1 \right\} \tag{42}$$

Remark

This criterion is know

- As the **Least Absolute Shrinkage and Selection Operator** (LASSO)
- This norm l_1 induces sparsity in the weight terms.

$$\left\Vert w \right\Vert_1 = \sum_{i=1}^{d} |w_i|$$

How?

- For example, $\left\Vert [1,0]^T \right\Vert_2 = \left\Vert [1/\sqrt{2}, 1/\sqrt{2}]^T \right\Vert_2 = 1$.
- In the other case, $\left\Vert [1,0]^T \right\Vert_1 = 1 < \left\Vert [1/\sqrt{2}, 1/\sqrt{2}]^T \right\Vert_1 = \sqrt{2}$.

Cinvestav

Remark

This criterion is know

- As the **Least Absolute Shrinkage and Selection Operator** (LASSO)
- This norm l_1 induces sparsity in the weight terms.

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |w_i|$$

How?

- For example, $\left\| [1, 0]^T \right\|_2 = \left\| [1/\sqrt{2}, 1/\sqrt{2}]^T \right\|_2 = 1.$
- In the other case, $\left\| [1, 0]^T \right\|_1 = 1 < \left\| [1/\sqrt{2}, 1/\sqrt{2}]^T \right\|_1 = \sqrt{2}.$

Remark

This criterion is know

- As the **Least Absolute Shrinkage and Selection Operator** (LASSO)
- This norm l_1 induces sparsity in the weight terms.

$$\|\boldsymbol{w}\|_1 = \sum_{i=1}^{d} |w_i|$$

How?

- For example, $\left\| [1, 0]^T \right\|_2 = \left\| [1/\sqrt{2}, 1/\sqrt{2}]^T \right\|_2 = 1.$
- In the other case, $\left\| [1, 0]^T \right\|_1 = 1 < \left\| [1/\sqrt{2}, 1/\sqrt{2}]^T \right\|_1 = \sqrt{2}.$

Remark

This criterion is know

- As the **Least Absolute Shrinkage and Selection Operator** (LASSO)
- This norm l_1 induces sparsity in the weight terms.

$$\left\|\boldsymbol{w}\right\|_1 = \sum_{i=1}^{d} |w_i|$$

How?

- For example, $\left\|[1,0]^T\right\|_2 = \left\|[1/\sqrt{2}, 1/\sqrt{2}]^T\right\|_2 = 1$.
- In the other case, $\left\|[1,0]^T\right\|_1 = 1 < \left\|[1/\sqrt{2}, 1/\sqrt{2}]^T\right\|_1 = \sqrt{2}$.

An example

What if $\boldsymbol{X}$ is a orthogonal matrix

In this case $\boldsymbol{X}^T\boldsymbol{X} = I$

Thus

An example

What if $\boldsymbol{X}$ is a orthogonal matrix

In this case $\boldsymbol{X}^T\boldsymbol{X} = I$

Thus

$$
\begin{aligned}
\widehat{\boldsymbol{w}} =& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + 2\sigma^2\alpha \left\| \boldsymbol{w} \right\|_1 \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right)^T \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right) + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\}
\end{aligned}
$$

An example

What if $\boldsymbol{X}$ is a orthogonal matrix

In this case $\boldsymbol{X}^T\boldsymbol{X} = I$

Thus

$$
\begin{aligned}
\widehat{\boldsymbol{w}} =& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + 2\sigma^2\alpha \|\boldsymbol{w}\|_1 \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\}
\end{aligned}
$$

An example

What if $\boldsymbol{X}$ is a orthogonal matrix

In this case $\boldsymbol{X}^T\boldsymbol{X} = I$

Thus

$$
\begin{aligned}
\widehat{\boldsymbol{w}} =& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \|\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}\|_2^2 + 2\sigma^2\alpha \|\boldsymbol{w}\|_1 \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y})^T (\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y}) + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\}
\end{aligned}
$$

An example

What if $\boldsymbol{X}$ is a orthogonal matrix

In this case $\boldsymbol{X}^T\boldsymbol{X} = I$

Thus

$$
\begin{aligned}
\widehat{\boldsymbol{w}} =& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left\| \boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right\|_2^2 + 2\sigma^2\alpha \left\| \boldsymbol{w} \right\|_1 \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right)^T \left(\boldsymbol{X}\boldsymbol{w} - \boldsymbol{y} \right) + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\} \\
=& \underset{\boldsymbol{w}}{\text{argmin}} \left\{ \boldsymbol{w}^T\boldsymbol{w} - 2\boldsymbol{w}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{y}^T\boldsymbol{y} + 2\sigma^2\alpha \sum_{i=1}^{d} |w_i| \right\}
\end{aligned}
$$

We can solve this last part as follow

We can group for each w_i

$$w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha\left|w_i\right| + y_i^2 \quad (43)$$

If we can minimize each group we will be able to get the solution

$$\widehat{w}_i = \underset{w_i}{\text{argmin}}\left\{w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha\left|w_i\right|\right\} \quad (44)$$

We have two cases

- $w_i > 0$
- $w_i < 0$

Cinvestav

We can solve this last part as follow

We can group for each w_i

$$w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha\left|w_i\right| + y_i^2 \tag{43}$$

If we can minimize each group we will be able to get the solution

$$\widehat{w}_i = \underset{w_i}{\text{argmin}}\left\{w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha\left|w_i\right|\right\} \tag{44}$$

We have two cases

- $w_i > 0$
- $w_i < 0$

We can solve this last part as follow

We can group for each w_i

$$w_i^2 - 2w_i \left(\boldsymbol{X}^T \boldsymbol{y}\right)_i + 2\sigma^2\alpha \left|w_i\right| + y_i^2 \tag{43}$$

If we can minimize each group we will be able to get the solution

$$\widehat{w}_i = \underset{w_i}{\text{argmin}} \left\{ w_i^2 - 2w_i \left(\boldsymbol{X}^T \boldsymbol{y}\right)_i + 2\sigma^2\alpha \left|w_i\right| \right\} \tag{44}$$

We have two cases

- $w_i > 0$
- $w_i < 0$

We can solve this last part as follow

We can group for each w_i

$$w_i^2 - 2w_i \left(\boldsymbol{X}^T \boldsymbol{y}\right)_i + 2\sigma^2 \alpha |w_i| + y_i^2 \tag{43}$$

If we can minimize each group we will be able to get the solution

$$\widehat{w}_i = \underset{w_i}{\text{argmin}} \left\{ w_i^2 - 2w_i \left(\boldsymbol{X}^T \boldsymbol{y}\right)_i + 2\sigma^2 \alpha |w_i| \right\} \tag{44}$$

We have two cases

- $w_i > 0$
- $w_i < 0$

If $w_i > 0$

We then derive with respect to w_i

$$\frac{\partial\left(w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha_i w_i\right)}{\partial w_i} = 2w_i - 2\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha$$

We have then

$$\widehat{w}_i = \left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - \sigma^2\alpha \tag{45}$$

If $w_i > 0$

We then derive with respect to w_i

$$\frac{\partial\left(w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha_i w_i\right)}{\partial w_i} = 2w_i - 2\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + 2\sigma^2\alpha$$

We have then

$$\widehat{w}_i = \left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - \sigma^2\alpha \tag{45}$$

If $w_i < 0$

We then derive with respect to w_i

$$\frac{\partial\left(w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - 2\sigma^2\alpha_i w_i\right)}{\partial w_i} = 2w_i - 2\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - 2\sigma^2\alpha$$

We have then

$$\widehat{w}_i = \left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + \sigma^2\alpha \quad (46)$$

Cinvestav

If $w_i < 0$

We then derive with respect to w_i

$$\frac{\partial\left(w_i^2 - 2w_i\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - 2\sigma^2\alpha_i w_i\right)}{\partial w_i} = 2w_i - 2\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i - 2\sigma^2\alpha$$

We have then

$$\widehat{w}_i = \left(\boldsymbol{X}^T\boldsymbol{y}\right)_i + \sigma^2\alpha \tag{46}$$

The value of $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i$

Ww have that

We have that:

- if $w_i > 0$ then $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i > \sigma^2\alpha$
- if $w_i < 0$ then $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i < -\sigma^2\alpha$

Cinvestav

We can put all this together

A compact Version

$$\widehat{w}_i = \mathsf{sgn}\left(\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right)\left(\left|\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right| - \sigma^2\alpha\right)_+ \tag{47}$$

With

$$(a)_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases}$$

- Where $(a)_+$ is the sign function.

This rule is know as the

- The soft threshold!!!

We can put all this together

A compact Version

$$\widehat{w}_i = \text{sgn}\left(\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right)\left(\left|\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right| - \sigma^2\alpha\right)_+ \tag{47}$$

With

$$(a)_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases}$$

- Where $(a)_+$ is the sign function.

This rule is know as the

- The soft threshold!!!

We can put all this together

A compact Version

$$\widehat{w}_i = \text{sgn}\left(\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right)\left(\left|\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i\right| - \sigma^2\alpha\right)_+ \tag{47}$$

With

$$(a)_+ = \begin{cases} a & \text{if } a \geq 0 \\ 0 & \text{if } a < 0 \end{cases}$$

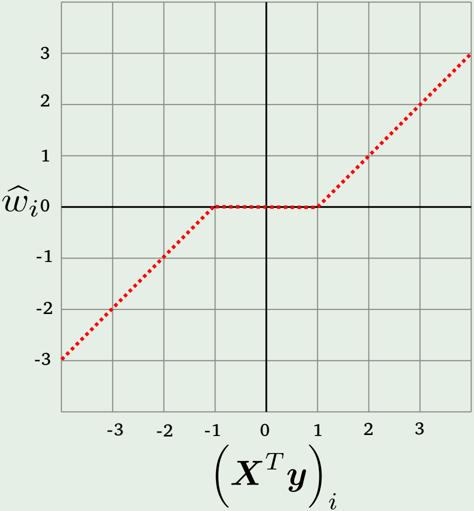
- Where $(a)_+$ is the sign function.

This rule is know as the

- The soft threshold!!!

Example

We have the doted $\cdots$ line as the soft threshold



Cinvestav

Outline

Cinvestav

Now, we need an estimate of each w_i

Given that each w_i has a zero-mean Gaussian prior

$$p\left(w_i|\tau_i\right) = \mathcal{N}\left(w_i|0, \tau_i\right) \tag{48}$$

Where τ_i has the following exponential hyper-prior

$$p\left(\tau_i|\gamma\right) = \frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\} \text{ for } \tau_i \geq 0 \tag{49}$$

This is a though property - so we take it by heart

$$p\left(w_i|\gamma\right) = \int_0^\infty p\left(w_i|\tau_i\right)p\left(\tau_i|\gamma\right)d\tau_i = \frac{\sqrt{\gamma}}{2}\exp\left\{-\sqrt{\gamma}\left|w_i\right|\right\} \tag{50}$$

Cinvestav

Now, we need an estimate of each w_i

Given that each w_i has a zero-mean Gaussian prior

$$p\left(w_i|\tau_i\right) = \mathcal{N}\left(w_i|0,\tau_i\right) \tag{48}$$

Where τ_i has the following exponential hyper-prior

$$p\left(\tau_i|\gamma\right) = \frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\} \text{ for } \tau_i \geq 0 \tag{49}$$

This is a though property - so we take it by heart

$$p\left(w_i|\gamma\right) = \int_0^{\infty} p\left(w_i|\tau_i\right)p\left(\tau_i|\gamma\right)d\tau_i = \frac{\sqrt{\gamma}}{2}\exp\left\{-\sqrt{\gamma}\left|w_i\right|\right\} \tag{50}$$

Cinvestav

Now, we need an estimate of each w_i

Given that each w_i has a zero-mean Gaussian prior

$$p\left(w_i|\tau_i\right) = \mathcal{N}\left(w_i|0,\tau_i\right) \tag{48}$$

Where τ_i has the following exponential hyper-prior

$$p\left(\tau_i|\gamma\right) = \frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\} \text{ for } \tau_i \geq 0 \tag{49}$$

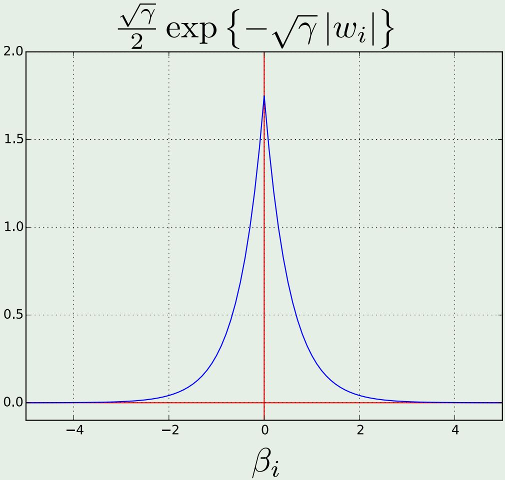
This is a though property - so we take it by heart

$$p\left(w_i|\gamma\right) = \int_0^\infty p\left(w_i|\tau_i\right)p\left(\tau_i|\gamma\right)d\tau_i = \frac{\sqrt{\gamma}}{2}\exp\left\{-\sqrt{\gamma}\left|w_i\right|\right\} \tag{50}$$

Cinvestav

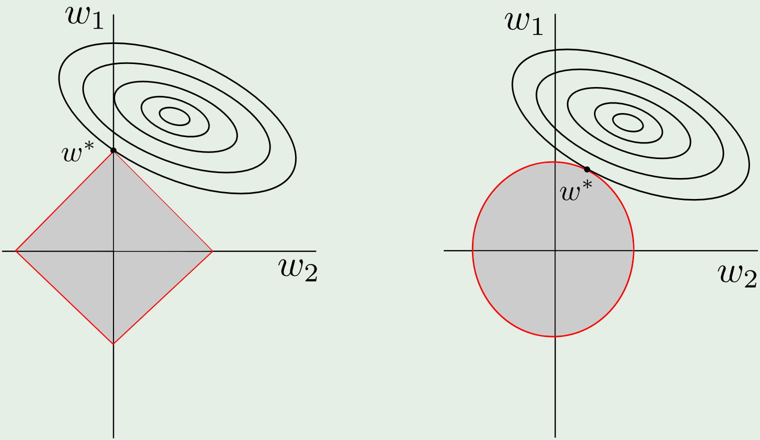
Example

The double exponential



This is equivalent to the use of the L_1-norm for regularization

L_1(Left) is better for sparsity promotion than L_2(Right)



Outline

The EM trick

How do we do this?

This is done by regarding $\tau = [\tau_1, ..., \tau_d]$ as the hidden/missing data

Then, if we could observe τ, complete log-posterior $\log p(w, \sigma^2 | y, \tau)$ can be easily calculated

$$p\left(w, \sigma^2 | y, \tau\right) \propto p\left(y | w, \sigma^2\right) p\left(w | \tau\right) p\left(\sigma^2\right) \quad (51)$$

The EM trick

How do we do this?

This is done by regarding $\tau = [\tau_1, ..., \tau_d]$ as the hidden/missing data

Then, if we could observe τ, complete log-posterior $\log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right)$ can be easily calculated

$$p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) \propto p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) p\left(\boldsymbol{w} | \tau\right) p\left(\sigma^2\right) \tag{51}$$

Where

- $p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) \sim N\left(\boldsymbol{X}\boldsymbol{w}, \sigma^2 I\right)$
- $p(\boldsymbol{w}|0, \tau) \sim \prod_{i=1}^{d} N\left(w_i | 0, \tau_i\right) = N\left(0, diag\left(\tau_1^{-1}, ..., \tau_d^{-1}\right)\right)$

The EM trick

How do we do this?

This is done by regarding $\tau = [\tau_1, ..., \tau_d]$ as the hidden/missing data

Then, if we could observe τ, complete log-posterior $\log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right)$ can be easily calculated

$$p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) \propto p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) p\left(\boldsymbol{w} | \tau\right) p\left(\sigma^2\right) \tag{51}$$

Where

- $p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) \sim \mathcal{N}\left(\boldsymbol{X}\boldsymbol{w}, \sigma^2 I\right)$
- $p\left(\boldsymbol{w} | 0, \tau\right) \sim \prod_{i=1}^{k} \mathcal{N}\left(w_i | 0, \tau_i\right) = \mathcal{N}\left(0, diag\left(\tau_1^{-1}, ..., \tau_d^{-1}\right)\right)$

What about $p\left(\sigma^2\right)$?

We select

$p\left(\sigma^2\right)$ as a constant

However

We can adopt a conjugate inverse Gamma prior for σ^2, but for large number of samples the prior on the estimate of $\sigma^2$is very small.

In the constant case

We can use the MAP idea, however we have hidden parameters so we resort to the EM

Cinvestav

What about $p\left(\sigma^2\right)$?

We select

$p\left(\sigma^2\right)$ as a constant

However

We can adopt a conjugate inverse Gamma prior for σ^2, but for large number of samples the prior on the estimate of $\sigma^2$is very small.

In the constant case

We can use the MAP idea, however we have hidden parameters so we resort to the EM

Cinvestav

What about $p(\sigma^2)$?

We select

$p(\sigma^2)$ as a constant

However

We can adopt a conjugate inverse Gamma prior for σ^2, but for large number of samples the prior on the estimate of $\sigma^2$is very small.

In the constant case

We can use the MAP idea, however we have hidden parameters so we resort to the EM

E-step

Computes the expected value of the complete log-posterior

$$Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = \int \log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \quad (52)$$

M-step

Updates the parameter estimates by maximizing the Q-function

$$\left(\widehat{\boldsymbol{w}}_{(t+1)}, \widehat{\sigma^2}_{(t+1)}\right) = \underset{\boldsymbol{w}, \sigma^2}{\text{argmax}} Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) \quad (53)$$

Remark

First

The EM algorithm converges to a local maximum of the a posteriori probability density function

$$p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}\right) \propto p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) p\left(\boldsymbol{w} | \gamma\right) \tag{54}$$

Without using the marginal prior $p\left(w|\gamma\right)$ which is not Gaussian

Instead we use a conditional Gaussian prior $p\left(w|\gamma\right)$

Remark

First

The EM algorithm converges to a local maximum of the a posteriori probability density function

$$p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}\right) \propto p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) p\left(\boldsymbol{w} | \gamma\right) \quad (54)$$

Without using the marginal prior $p\left(\boldsymbol{w} | \gamma\right)$ which is not Gaussian

Instead we use a conditional Gaussian prior $p\left(\boldsymbol{w} | \gamma\right)$

The Final Model

We have

$$p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right)=\mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma^2 I\right)$$

$$p\left(\sigma^2\right)\propto \text{"}constant\text{"}$$

$$p\left(w|\tau\right)=\prod_{i=1}^{d}\mathcal{N}\left(w_i|0,\tau_i\right)=\mathcal{N}\left(\boldsymbol{w}|0,\left(\Upsilon\left(\tau\right)\right)^{-1}\right)$$

$$p\left(\tau|\gamma\right)=\left(\frac{\gamma}{2}\right)^{d}\prod_{i=1}^{d}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}$$

With $\Upsilon\left(\tau\right)=diag\left(\tau_1^{-1},...,\tau_d^{-1}\right)$ is the diagonal matrix with the inverse variances of all the w_i's.

The Final Model

We have

$$p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma^2 I\right)$$

$$p\left(\sigma^2\right) \propto \text{"}constant\text{"}$$

$$p\left(w|\tau\right) = \prod_{i=1}^{d} \mathcal{N}\left(w_i|0,\tau_i\right) = \mathcal{N}\left(\boldsymbol{w}|0,\left(\Upsilon\left(\tau\right)\right)^{-1}\right)$$

$$p\left(\tau|\gamma\right) = \left(\frac{\gamma}{2}\right)^{d} \prod_{i=1}^{d} \exp\left\{-\frac{\gamma}{2}\tau_i\right\}$$

With $\Upsilon\left(\tau\right) = diag\left(\tau_1^{-1},...,\tau_d^{-1}\right)$ is the diagonal matrix with the inverse variances of all the w_i's.

The Final Model

We have

$$p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right)=\mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma^2 I\right)$$

$$p\left(\sigma^2\right)\propto \text{"}constant\text{"}$$

$$p\left(\boldsymbol{w}|\tau\right)=\prod_{i=1}^{d}\mathcal{N}\left(w_i|0,\tau_i\right)=\mathcal{N}\left(\boldsymbol{w}|0,\left(\Upsilon\left(\tau\right)\right)^{-1}\right)$$

$$p\left(\tau|\gamma\right)=\left(\frac{\gamma}{2}\right)^{d}\prod_{i=1}^{d}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}$$

With $\Upsilon\left(\tau\right)=diag\left(\tau_1^{-1},...,\tau_d^{-1}\right)$ is the diagonal matrix with the inverse variances of all the w_i's.

The Final Model

We have

$$p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right) = \mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma^2 I\right)$$

$$p\left(\sigma^2\right) \propto \text{"}constant\text{"}$$

$$p\left(\boldsymbol{w}|\tau\right) = \prod_{i=1}^{d} \mathcal{N}\left(w_i|0,\tau_i\right) = \mathcal{N}\left(\boldsymbol{w}|0,\left(\Upsilon\left(\tau\right)\right)^{-1}\right)$$

$$p\left(\tau|\gamma\right) = \left(\frac{\gamma}{2}\right)^{d} \prod_{i=1}^{d} \exp\left\{-\frac{\gamma}{2}\tau_i\right\}$$

With $\Upsilon\left(\tau\right) = diag\left(\tau_1^{-1},...,\tau_d^{-1}\right)$ is the diagonal matrix with the inverse variances of all the w_i's.

The Final Model

We have

$$p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right)=\mathcal{N}\left(\boldsymbol{X}\boldsymbol{w},\sigma^2 I\right)$$

$$p\left(\sigma^2\right)\propto \text{"}constant\text{"}$$

$$p\left(\boldsymbol{w}|\tau\right)=\prod_{i=1}^{d}\mathcal{N}\left(w_i|0,\tau_i\right)=\mathcal{N}\left(\boldsymbol{w}|0,\left(\Upsilon\left(\tau\right)\right)^{-1}\right)$$

$$p\left(\tau|\gamma\right)=\left(\frac{\gamma}{2}\right)^{d}\prod_{i=1}^{d}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}$$

With $\Upsilon\left(\tau\right)=diag\left(\tau_1^{-1},...,\tau_d^{-1}\right)$ is the diagonal matrix with the inverse variances of all the w_i's.

Now, we find the Q function

First

$$\log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) \propto \log p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) + \log p\left(\boldsymbol{w} | \tau\right)$$

$$\propto -n \log \sigma^2 - \frac{\|y - Xw\|_2^2}{\sigma^2} - w^T \Upsilon(\tau) w$$

How can we get this?

Remember

$$\mathcal{N}(y|\mu, \Sigma) = \frac{1}{(2\pi)^{\frac{k}{2}} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right\} \quad (55)$$

Volunteers?

Please to the blackboard.

Now, we find the Q function

First

$$\log p\left(\boldsymbol{w},\sigma^2|\boldsymbol{y},\tau\right) \propto \log p\left(\boldsymbol{y}|\boldsymbol{w},\sigma^2\right) + \log p\left(\boldsymbol{w}|\tau\right)$$
$$\propto -n\log\sigma^2 - \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T\Upsilon\left(\tau\right)\boldsymbol{w}$$

How can we get this?

Remember

$$\mathcal{N}\left(\boldsymbol{y}|\boldsymbol{\mu},\boldsymbol{\Sigma}\right) = \frac{1}{(2\pi)^{\frac{k}{2}}\left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}}\exp\left\{-\frac{1}{2}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)^T\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}-\boldsymbol{\mu}\right)\right\} \tag{55}$$

Volunteers?

Please to the blackboard.

Cinvestav

Now, we find the Q function

First

$$
\begin{aligned}
\log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) &\propto \log p\left(\boldsymbol{y} | \boldsymbol{w}, \sigma^2\right) + \log p\left(\boldsymbol{w} | \tau\right) \\
&\propto -n \log \sigma^2 - \frac{\left\| \boldsymbol{y} - \boldsymbol{X}\boldsymbol{w} \right\|_2^2}{\sigma^2} - \boldsymbol{w}^T \Upsilon\left(\tau\right) \boldsymbol{w}
\end{aligned}
$$

How can we get this?

Remember

$$
\mathcal{N}\left(\boldsymbol{y} | \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) = \frac{1}{\left(2\pi\right)^{\frac{k}{2}} \left|\boldsymbol{\Sigma}\right|^{\frac{1}{2}}} \exp\left\{ -\frac{1}{2} \left(\boldsymbol{y} - \boldsymbol{\mu}\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{y} - \boldsymbol{\mu}\right) \right\} \quad (55)
$$

Volunteers?

Please to the blackboard.

Thus

Second

Did you notice that the term $\boldsymbol{w}^T \Upsilon(\tau)\, \boldsymbol{w}$ is linear with respect to $\Upsilon(\tau)$ and the other terms do not depend on τ?

Thus

Second

Did you notice that the term $\boldsymbol{w}^T \Upsilon(\tau)\, \boldsymbol{w}$ is linear with respect to $\Upsilon(\tau)$ and the other terms do not depend on τ?

Thus, the E-step is reduced to the computation of $\Upsilon(\tau)$

$$\boldsymbol{V}_{(t)} = E\left(\Upsilon(\tau)\,|\boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right)$$

$$= diag\left(E\left[\tau_1^{-1}|\boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right], \ldots, E\left[\tau_d^{-1}|\boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right]\right)$$

Cinvestav

Thus

Second

Did you notice that the term $\boldsymbol{w}^T \Upsilon(\tau) \boldsymbol{w}$ is linear with respect to $\Upsilon(\tau)$ and the other terms do not depend on τ?

Thus, the E-step is reduced to the computation of $\Upsilon(\tau)$

$$\begin{aligned}\boldsymbol{V}_{(t)} &= E\left(\Upsilon(\tau) \mid \boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) \\ &= diag\left(E\left[\tau_1^{-1} | \boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right], ..., E\left[\tau_d^{-1} | \boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right]\right)\end{aligned}$$

Now

What do we need to calculate each of this expectations?

$$p\left(\tau_i | \boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = p\left(\tau_i | \boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{i,(t)}\right) \quad (56)$$

Then

Now

What do we need to calculate each of this expectations?

$$p\left(\tau_i|\boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = p\left(\tau_i|\boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{i,(t)}\right) \tag{56}$$

Then

$$p\left(\tau_i|\boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{i,(t)}\right) = \frac{p\left(\tau_i, \widehat{w}_{i,(t)}|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right)}{p\left(\widehat{w}_{i,(t)}|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right)}$$

$$\propto p\left(\widehat{w}_{i,(t)}|\tau_i, \boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right) p\left(\tau_i|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right)$$

$$= p\left(\widehat{w}_{i,(t)}|\tau_i\right) p\left(\tau_i\right)$$

Now

What do we need to calculate each of this expectations?

$$p\left(\tau_i|\boldsymbol{y},\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)}\right)=p\left(\tau_i|\boldsymbol{y},\widehat{w}_{i,(t)},\widehat{\sigma^2}_{i,(t)}\right) \tag{56}$$

Then

$$\begin{aligned}
p\left(\tau_i|\boldsymbol{y},\widehat{w}_{i,(t)},\widehat{\sigma^2}_{i,(t)}\right) &= \frac{p\left(\tau_i,\widehat{w}_{i,(t)}|\boldsymbol{y},\widehat{\sigma^2}_{i,(t)}\right)}{p\left(\widehat{w}_{i,(t)}|\boldsymbol{y},\widehat{\sigma^2}_{i,(t)}\right)}\\
&\propto p\left(\widehat{w}_{i,(t)}|\tau_i,\boldsymbol{y},\widehat{\sigma^2}_{i,(t)}\right)p\left(\tau_i|\boldsymbol{y},\widehat{\sigma^2}_{i,(t)}\right)\\
&= p\left(\widehat{w}_{i,(t)}|\tau_i\right)p\left(\tau_i\right)
\end{aligned}$$

Now

What do we need to calculate each of this expectations?

$$p\left(\tau_i|\boldsymbol{y}, \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = p\left(\tau_i|\boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{i,(t)}\right) \quad (56)$$

Then

$$\begin{aligned}
p\left(\tau_i|\boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{i,(t)}\right) &= \frac{p\left(\tau_i, \widehat{w}_{i,(t)}|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right)}{p\left(\widehat{w}_{i,(t)}|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right)} \\
&\propto p\left(\widehat{w}_{i,(t)}|\tau_i, \boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right) p\left(\tau_i|\boldsymbol{y}, \widehat{\sigma^2}_{i,(t)}\right) \\
&= p\left(\widehat{w}_{i,(t)}|\tau_i\right) p\left(\tau_i\right)
\end{aligned}$$

Here is the interesting part

We have the following probability density

$$p\left(\tau_i|\boldsymbol{y},\widehat{\beta}_{i,(t)},\widehat{\sigma^2}_{i,(t)}\right)=\frac{\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}}{\int_0^{\infty}\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \tag{57}$$

Then

$$E\left[\tau_i^{-1}|\boldsymbol{y},\widehat{w}_{i,(t)},\widehat{\sigma^2}_{(t)}\right]=\frac{\int_0^{\infty}\frac{1}{\tau_i}\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i}{\int_0^{\infty}\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \tag{58}$$

Now

I leave to you to prove that (It can come in the test)

Here is the interesting part

We have the following probability density

$$p\left(\tau_i|\boldsymbol{y},\widehat{\beta}_{i,(t)},\widehat{\sigma^2}_{i,(t)}\right)=\frac{\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}}{\int_0^\infty\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \tag{57}$$

Then

$$E\left[\tau_i^{-1}|\boldsymbol{y},\widehat{w}_{i,(t)},\widehat{\sigma^2}_{(t)}\right]=\frac{\int_0^\infty\frac{1}{\tau_i}\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i}{\int_0^\infty\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \tag{58}$$

Now

I leave to you to prove that (It can come in the test)

Here is the interesting part

We have the following probability density

$$p\left(\tau_i|\boldsymbol{y},\widehat{\beta}_{i,(t)},\widehat{\sigma^2}_{i,(t)}\right)=\frac{\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}}{\int_0^\infty\mathcal{N}\left(\beta_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \quad (57)$$

Then

$$E\left[\tau_i^{-1}|\boldsymbol{y},\widehat{w}_{i,(t)},\widehat{\sigma^2}_{(t)}\right]=\frac{\int_0^\infty\frac{1}{\tau_i}\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i}{\int_0^\infty\mathcal{N}\left(w_{i,(t)}|0,\tau_i\right)\frac{\gamma}{2}\exp\left\{-\frac{\gamma}{2}\tau_i\right\}d\tau_i} \quad (58)$$

Now

I leave to you to prove that (It can come in the test)

Here is the interesting part

Thus

$$E\left[\tau_i^{-1} | \boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{(t)}\right] = \frac{\gamma}{\left|\widehat{w}_{i,(t)}\right|} \tag{59}$$

Finally

$$\boldsymbol{V}_{(t)} = \gamma diag\left(\left|\widehat{w}_{1,(t)}\right|^{-1}, \ldots, \left|\widehat{w}_{d,(t)}\right|^{-1}\right) \tag{60}$$

Here is the interesting part

Thus

$$E\left[\tau_i^{-1}|\boldsymbol{y}, \widehat{w}_{i,(t)}, \widehat{\sigma^2}_{(t)}\right] = \frac{\gamma}{\left|\widehat{w}_{i,(t)}\right|} \tag{59}$$

Finally

$$\boldsymbol{V}_{(t)} = \gamma diag\left(\left|\widehat{w}_{1,(t)}\right|^{-1}, ..., \left|\widehat{w}_{d,(t)}\right|^{-1}\right) \tag{60}$$

The Final Q function

Something Notable

$$
\begin{aligned}
Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) &= \int \log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= \int \left[-n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \Upsilon(\tau) \boldsymbol{w}\right] p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \left[\int \Upsilon(\tau) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau\right] \boldsymbol{w} \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w}
\end{aligned}
$$

Cinvestav

The Final Q function

Something Notable

$$
\begin{aligned}
Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) &= \int \log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= \int \left[-n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \Upsilon(\tau)\, \boldsymbol{w}\right] p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \left[\int \Upsilon(\tau)\, p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau\right] \boldsymbol{w} \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T V_{(t)} \boldsymbol{w}
\end{aligned}
$$

The Final Q function

Something Notable

$$
\begin{aligned}
Q\left(\boldsymbol{w},\sigma^2|\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)}\right) &= \int \log p\left(\boldsymbol{w},\sigma^2|\boldsymbol{y},\tau\right) p\left(\tau|\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)},\boldsymbol{y}\right) d\tau \\
&= \int \left[-n\log\sigma^2 - \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T\Upsilon(\tau)\,\boldsymbol{w}\right] p\left(\tau|\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)},\boldsymbol{y}\right) d\tau \\
&= -n\log\sigma^2 - \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T\left[\int \Upsilon(\tau)\,p\left(\tau|\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)},\boldsymbol{y}\right) d\tau\right]\boldsymbol{w} \\
&= -n\log\sigma^2 - \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T V_{(t)}\boldsymbol{w}
\end{aligned}
$$

The Final Q function

Something Notable

$$
\begin{aligned}
Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) &= \int \log p\left(\boldsymbol{w}, \sigma^2 | \boldsymbol{y}, \tau\right) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= \int \left[-n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \Upsilon(\tau) \boldsymbol{w}\right] p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \left[\int \Upsilon(\tau) p\left(\tau | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}, \boldsymbol{y}\right) d\tau\right] \boldsymbol{w} \\
&= -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w}
\end{aligned}
$$

Cinvestav

Finally, the M-step

First

$$\widehat{\sigma^2}_{(t+1)} = \underset{\sigma^2}{\text{argmax}} \left\{ -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} \right\}$$

$$= \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{n}$$

Second

$$\widehat{\boldsymbol{w}}_{(t+1)} = \underset{\boldsymbol{w}}{\text{argmax}} \left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w} \right\}$$

$$= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

This also I leave to you

It can come in the test.

Finally, the M-step

First

$$
\begin{aligned}
\widehat{\sigma^2}_{(t+1)} &= \underset{\sigma^2}{\operatorname{argmax}} \left\{ -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} \right\} \\
&= \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{n}
\end{aligned}
$$

Second

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \underset{\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w} \right\} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}
\end{aligned}
$$

This also I leave to you

It can come in the test.

Finally, the M-step

First

$$\widehat{\sigma^2}_{(t+1)} = \underset{\sigma^2}{\operatorname{argmax}} \left\{ -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} \right\}$$

$$= \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{n}$$

Second

$$\widehat{\boldsymbol{w}}_{(t+1)} = \underset{\boldsymbol{w}}{\operatorname{argmax}} \left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w} \right\}$$

$$= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

This also I leave to you

It can come in the test.

Finally, the M-step

First

$$\widehat{\sigma^2}_{(t+1)} = \underset{\sigma^2}{\text{argmax}} \left\{ -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} \right\}$$
$$= \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{n}$$

Second

$$\widehat{\boldsymbol{w}}_{(t+1)} = \underset{\boldsymbol{w}}{\text{argmax}} \left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w} \right\}$$
$$= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

This also I leave to you

It can come in the test.

Finally, the M-step

First

$$\widehat{\sigma^2}_{(t+1)} = \underset{\sigma^2}{\text{argmax}} \left\{ -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} \right\}$$
$$= \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{n}$$

Second

$$\widehat{\boldsymbol{w}}_{(t+1)} = \underset{\boldsymbol{w}}{\text{argmax}} \left\{ -\frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w} \right\}$$
$$= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X} \right)^{-1} \boldsymbol{X}^T \boldsymbol{y}$$

This also I leave to you

It can come in the test.

Outline

However

We need to deal in some way with the γ term

It controls the degree of spareness!!!

We can do assuming a Jeffrey's Prior

J. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1980.

We use instead

$$p(\tau) \propto \frac{1}{\tau} \tag{61}$$

However

We need to deal in some way with the γ term

It controls the degree of spareness!!!

We can do assuming a Jeffrey's Prior

J. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1980.

We use instead

$$p(\tau) \propto \frac{1}{\tau} \tag{61}$$

However

We need to deal in some way with the γ term

It controls the degree of spareness!!!

We can do assuming a Jeffrey's Prior

J. Berger, *Statistical Decision Theory and Bayesian Analysis*. New York: Springer-Verlag, 1980.

We use instead

$$p(\tau) \propto \frac{1}{\tau} \tag{61}$$

Properties of the Jeffrey's Prior

Important

This prior expresses ignorance with respect to scale and is parameter free

Why scale invariant

Imagine, we change the scale of τ by $\tau' = K\tau$ where K is a constant expressing that change

Thus, we have that

$$p(\tau') = \frac{1}{\tau'} = \frac{1}{k\tau} \propto \frac{1}{\tau} \tag{62}$$

Properties of the Jeffrey's Prior

Important

This prior expresses ignorance with respect to scale and is parameter free

Why scale invariant

Imagine, we change the scale of τ by $\tau' = K\tau$ where K is a constant expressing that change

Thus, we have that

$$p\left(\tau'\right) = \frac{1}{\tau'} = \frac{1}{k\tau} \propto \frac{1}{\tau} \tag{62}$$

Properties of the Jeffrey's Prior

Important

This prior expresses ignorance with respect to scale and is parameter free

Why scale invariant

Imagine, we change the scale of τ by $\tau' = K\tau$ where K is a constant expressing that change

Thus, we have that

$$p\left(\tau'\right) = \frac{1}{\tau'} = \frac{1}{k\tau} \propto \frac{1}{\tau} \tag{62}$$

Cinvestav

However

Something Notable

This prior is known as an improper prior.

In addition

This prior does not leads to a Laplacian prior on w.

Nevertheless

This prior induces sparseness and good performance for the w.

Cinvestav

However

Something Notable

This prior is known as an improper prior.

In addition

This prior does not leads to a Laplacian prior on $\boldsymbol{w}$.

Nevertheless

This prior induces sparseness and good performance for the $\boldsymbol{w}$.

However

Something Notable

This prior is known as an improper prior.

In addition

This prior does not leads to a Laplacian prior on $\boldsymbol{w}$.

Nevertheless

This prior induces sparseness and good performance for the $\boldsymbol{w}$.

Introducing this prior into the equations

Matrix $\boldsymbol{V}_{(t)}$ is now

$$\boldsymbol{V}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|^{-2}, ..., \left|\widehat{w}_{d,(t)}\right|^{-2}\right) \tag{63}$$

Quite interesting!!!

We do not have the free γ parameter.

Introducing this prior into the equations

Matrix $\boldsymbol{V}_{(t)}$ is now

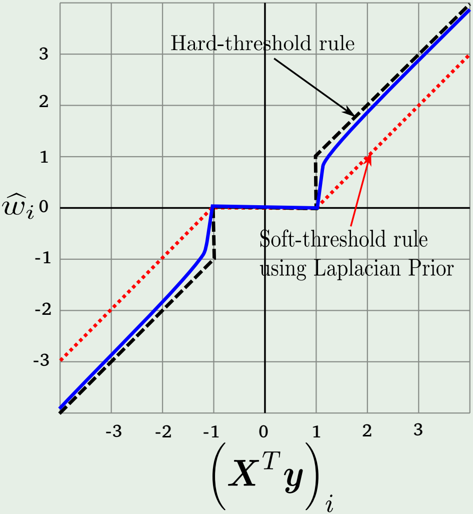
$$\boldsymbol{V}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|^{-2}, ..., \left|\widehat{w}_{d,(t)}\right|^{-2}\right) \quad (63)$$

Quite interesting!!!

We do not have the free γ parameter.

Here, we can see the new threshold

Blue solid line - estimation rule using EM and Jeffrey's Hyperprior



Observations

The new rule is between

- The soft threshold rule.
- The hard threshold rule.

Something Notable

With large values of $\left(X^T y\right)_i$ the new rule approaches the hard threshold.

Once $\left(X^T y\right)_i$ gets smaller

The estimate becomes progressively smaller approaching the behavior of the soft rule.

Observations

The new rule is between

- The soft threshold rule.
- The hard threshold rule.

Something Notable

With large values of $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i$ the new rule approaches the hard threshold.

Once $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i$ gets smaller

The estimate becomes progressively smaller approaching the behavior of the soft rule.

Cinvestav

Observations

The new rule is between

- The soft threshold rule.
- The hard threshold rule.

Something Notable

With large values of $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i$ the new rule approaches the hard threshold.

Once $\left(\boldsymbol{X}^T\boldsymbol{y}\right)_i$ gets smaller

The estimate becomes progressively smaller approaching the behavior of the soft rule.

Cinvestav

Finally, an implementation detail

Since several elements of $\widehat{\boldsymbol{w}}$ will go to zero

$\boldsymbol{V}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|^{-2}, ..., \left|\widehat{w}_{d,(t)}\right|^{-2}\right)$ will have several elements going to large numbers

Something Notable

if we define $\boldsymbol{U}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|, ..., \left|\widehat{w}_{d,(t)}\right|\right)$.

Then, we have that

$$\boldsymbol{V}_{(t)} = \boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}^{-1} \quad (64)$$

Finally, an implementation detail

Since several elements of $\widehat{\boldsymbol{w}}$ will go to zero

$\boldsymbol{V}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|^{-2}, ..., \left|\widehat{w}_{d,(t)}\right|^{-2}\right)$ will have several elements going to large numbers

Something Notable

if we define $\boldsymbol{U}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|, ..., \left|\widehat{w}_{d,(t)}\right|\right)$.

Then, we have that

$$\boldsymbol{V}_{(t)} = \boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}^{-1} \tag{64}$$

Finally, an implementation detail

Since several elements of $\widehat{\boldsymbol{w}}$ will go to zero

$\boldsymbol{V}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|^{-2}, ..., \left|\widehat{w}_{d,(t)}\right|^{-2}\right)$ will have several elements going to large numbers

Something Notable

if we define $\boldsymbol{U}_{(t)} = diag\left(\left|\widehat{w}_{1,(t)}\right|, ..., \left|\widehat{w}_{d,(t)}\right|\right)$.

Then, we have that

$$\boldsymbol{V}_{(t)} = \boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}^{-1} \tag{64}$$

Cinvestav

Thus

We have that

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + I \boldsymbol{X}^T \boldsymbol{X} I\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)} \boldsymbol{U}_{(t)}^{-1}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \boldsymbol{U}_{(t)} \left(\widehat{\sigma^2}_{(t+1)} I + \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)}\right)^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{y}
\end{aligned}
$$

Thus

We have that

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + I \boldsymbol{X}^T \boldsymbol{X} I\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)} \boldsymbol{U}_{(t)}^{-1}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \boldsymbol{U}_{(t)} \left(\widehat{\sigma^2}_{(t+1)} I + \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)}\right)^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{y}
\end{aligned}
$$

Thus

We have that

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{V}_{(t)} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}^{-1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}I\boldsymbol{U}_{(t)}^{-1} + I\boldsymbol{X}^T\boldsymbol{X}I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}I\boldsymbol{U}_{(t)}^{-1} + \boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{U}_{(t)}\boldsymbol{U}_{(t)}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= \boldsymbol{U}_{(t)}\left(\widehat{\sigma^2}_{(t+1)}I + \boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{U}_{(t)}\right)^{-1}\boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Thus

We have that

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{V}_{(t)} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}^{-1} + \boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}I\boldsymbol{U}_{(t)}^{-1} + I\boldsymbol{X}^T\boldsymbol{X}I\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \left(\widehat{\sigma^2}_{(t+1)}\boldsymbol{U}_{(t)}^{-1}I\boldsymbol{U}_{(t)}^{-1} + \boldsymbol{U}_{(t)}^{-1}\boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{U}_{(t)}\boldsymbol{U}_{(t)}^{-1}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}\\
&= \boldsymbol{U}_{(t)}\left(\widehat{\sigma^2}_{(t+1)}I + \boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{U}_{(t)}\right)^{-1}\boldsymbol{U}_{(t)}\boldsymbol{X}^T\boldsymbol{y}
\end{aligned}
$$

Thus

We have that

$$
\begin{aligned}
\widehat{\boldsymbol{w}}_{(t+1)} &= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{V}_{(t)} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + I \boldsymbol{X}^T \boldsymbol{X} I\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \left(\widehat{\sigma^2}_{(t+1)} \boldsymbol{U}_{(t)}^{-1} I \boldsymbol{U}_{(t)}^{-1} + \boldsymbol{U}_{(t)}^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)} \boldsymbol{U}_{(t)}^{-1}\right)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\
&= \boldsymbol{U}_{(t)} \left(\widehat{\sigma^2}_{(t+1)} I + \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{X} \boldsymbol{U}_{(t)}\right)^{-1} \boldsymbol{U}_{(t)} \boldsymbol{X}^T \boldsymbol{y}
\end{aligned}
$$

Advantages!!!

Quite Important

We avoid the inversion of the elements of $\widehat{\boldsymbol{w}}_{(t)}$.

We can avoid getting the inverse matrix

We simply solve the corresponding linear system whose dimension is only the number of nonzero elements in $U_{(t)}$. Why?

- Remember you want to maximize $Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w}$

Advantages!!!

Quite Important

We avoid the inversion of the elements of $\widehat{\boldsymbol{w}}_{(t)}$.

We can avoid getting the inverse matrix

We simply solve the corresponding linear system whose dimension is only the number of nonzero elements in $\boldsymbol{U}_{(t)}$. Why?

- Remember you want to maximize

$$Q\left(\boldsymbol{w},\sigma^2|\widehat{\boldsymbol{w}}_{(t)},\widehat{\sigma^2}_{(t)}\right) = -n\log\sigma^2 - \frac{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T\boldsymbol{V}_{(t)}\boldsymbol{w}$$

Advantages!!!

Quite Important

We avoid the inversion of the elements of $\widehat{\boldsymbol{w}}_{(t)}$.

We can avoid getting the inverse matrix

We simply solve the corresponding linear system whose dimension is only the number of nonzero elements in $\boldsymbol{U}_{(t)}$. Why?

- Remember you want to maximize $Q\left(\boldsymbol{w}, \sigma^2 | \widehat{\boldsymbol{w}}_{(t)}, \widehat{\sigma^2}_{(t)}\right) = -n \log \sigma^2 - \frac{\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{w}\|_2^2}{\sigma^2} - \boldsymbol{w}^T \boldsymbol{V}_{(t)} \boldsymbol{w}$

Cinvestav