

Introduction to Machine Learning - CS725
Instructor: Prof. Ganesh Ramakrishnan
Lecture 4 - Linear Regression - Bayesian Inference
and Regularization

Building on questions on Least Squares Linear Regression

- 1 Is there a probabilistic interpretation?
 - Gaussian Error, Maximum Likelihood Estimate
- 2 Addressing overfitting
 - Bayesian and Maximum A posteriori Estimates, Regularization
- 3 How to minimize the resultant and more complex error functions?
 - Level Curves and Surfaces, Gradient Vector, Directional Derivative, Gradient Descent Algorithm, Convexity, Necessary and Sufficient Conditions for Optimality

Recap: Bayesian Inference with Coin Tossing

Let $\mathcal{D} | H$ follow a distribution $Ber(p)$ (p is probability of heads) and p follow a distribution $Beta(p; \alpha, \beta) \sim \frac{p^{(\alpha-1)}(1-p)^{(\beta-1)}}{B(\alpha, \beta)}$,

- ① The Maximum Likelihood Estimate:

$$\hat{p} = \operatorname{argmax}_p {}^n C_h p^h (1-p)^{n-h} = \frac{h}{n}$$

- ② The Posterior Distribution: *Prior = Beta(α, β)*

$$\Pr(p | \mathcal{D}) = Beta(p; \alpha + h, \beta + n - h)$$

- ③ The Maximum a-Posterior (MAP) Estimate: The mode of the posterior distribution

$$\tilde{p} = \operatorname{argmax}_H \Pr(H | \mathcal{D}) = \operatorname{argmax}_p \Pr(p | \mathcal{D})$$

$$= \operatorname{argmax}_p Beta(p; \alpha + h, \beta + n - h) = \frac{\alpha + h - 1}{\alpha + \beta + n - 2}$$

$$E_{\text{posterior}}[p] = \int p Beta(p; \alpha + h, \beta + n - h) = \frac{\alpha + h}{\alpha + \beta + n}$$

Bayes Estimate

Intuition for Bayesian Linear Regression

- The Bayesian interpretation of probabilistic estimation is a logical extension that enables reasoning with uncertainty **but in the light of some background belief**
- **Bayesian linear regression**: A Bayesian alternative to **Maximum Likelihood** least squares regression
- Continue with Normally distributed errors
- Model the \mathbf{w} using a prior distribution and use the posterior over \mathbf{w} as the result
- **Intuitive Prior**: Components of \mathbf{w} should not become too large!

} A 3rd probabilistic form

2 forms of regularized Ridge regression:

$$\textcircled{1} \hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|_2^2 \quad \text{s.t. } \|\mathbf{w}\|_2^2 \leq \theta$$
$$\textcircled{2} \hat{\mathbf{w}} = \underset{\mathbf{w}}{\operatorname{argmin}} \|\phi\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_2^2$$

$\phi = m \times n$ matrix

Prior Distribution for w for Linear Regression

$$y = \mathbf{w}^T \phi(x) + \varepsilon$$

$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- We saw that when we try to maximize log-likelihood we end up with $\hat{\mathbf{w}}_{MLE} = (\phi^T \phi)^{-1} \phi^T y$

- We can use a Prior distribution on \mathbf{w} to avoid over-fitting

$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$

In absence of D_i ,
 $E(w_i) = 0$

Each component w_i is approximately bounded within $\pm \frac{3}{\sqrt{\lambda}}$. λ

is also called the precision of the Gaussian

$\frac{1}{\lambda}$ increasing \rightarrow

- Q1: How do deal with Bayesian Estimation for Gaussian distribution?

$$p(w_i | D) ?$$

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \dots x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\sigma_{MLE}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?

Conjugate Prior for (univariate) Gaussian

- We will temporarily generalize the discussion with x taking the place of ε and μ taking the place of w_i
- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \dots x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\sigma_{MLE}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?
- Answer: $\Pr(\mu | x_1 \dots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that $\mu_m = \dots$ and $\frac{1}{\sigma_m^2} = \dots$
- Helpful tip: Product of Gaussians is always a Gaussian

$$P_{\mu}(u) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(-\frac{(u-\mu_0)^2}{2\sigma_0^2}\right)$$

$$P_{\mu}(x_i | u) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - u)^2}{2\sigma^2}\right)$$

$$P_{\mu}(x_1, \dots, x_m | u) = \prod_{i=1}^m \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right) \exp\left(-\frac{(x_i - u)^2}{2\sigma^2}\right)$$

under iid assumption

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^m \exp\left(\sum_{i=1}^m -\frac{(x_i - u)^2}{2\sigma^2}\right) = L(u | D)$$

$$P_{\mu}(u | x_1, \dots, x_m) = \frac{P_{\mu}(x_1, \dots, x_m | u) P_{\mu}(u)}{P_{\mu}(x_1, \dots, x_m)} \propto P_{\mu}(x_1, \dots, x_m | u) P_{\mu}(u)$$

A normalizing factor
& we will bundle all terms independent of u
into normalization

Detailed derivation

$$\Pr(\mu) = \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-(\mu - \mu_0)^2}{2\sigma_0^2}\right)$$

$$\Pr(x_i|\mu; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(x_i - \mu)^2}{2\sigma^2}\right)$$

$$\Pr(\mathcal{D}|\mu) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2\right)$$

$$\Pr(\mu|\mathcal{D}) \propto \Pr(\mathcal{D}|\mu) \Pr(\mu) =$$

$$\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \frac{1}{\sqrt{2\pi\sigma_0^2}} \exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) \propto$$

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right)$$

My leap of faith: $P_i(\mu|x_1, \dots, x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$

Further ignore terms independent of μ

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$-\frac{\mu^2}{2\sigma_m^2} = -\frac{m\mu^2}{2\sigma^2} - \frac{\mu^2}{2\sigma_0^2} \Rightarrow \frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

Precision grows as data observed increases.

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2}\sum_{i=1}^m(x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2}(\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2}\left(\frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Rightarrow$$

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2} \left(\frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$$

$$\frac{-1}{2\sigma_m^2} = \underbrace{\frac{-1}{2\sigma^2} \sum_{i=1}^m (-2x_i)}_{-2\mu_m}$$

Matching coefficients of μ , we get

$$-\frac{1}{2\sigma_m^2} (-2\mu_m) = \frac{m}{\sigma^2} (-2\mu_0) - \frac{-2\mu_0}{2\sigma_0^2}$$

$$\frac{\mu_m}{\sigma_m^2} = \frac{\mu_0}{\sigma_0^2} + \frac{\sum x_i}{\sigma^2} = \frac{\mu_0}{\sigma_0^2} + \frac{\mu_{MLE} \approx m}{\sigma^2}$$

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2} \left(\frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$$

Matching coefficients of μ , we get

$$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu \left(\frac{2\sum_{i=1}^m x_i}{2\sigma^2} + \frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow$$

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2} \left(\frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma^2} + \frac{m}{\sigma_0^2} \quad \left[\sigma_m^2 \propto m(\sigma_0^2, \frac{\sigma^2}{m})\right]$$

Matching coefficients of μ , we get

$$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu \left(\frac{2\sum_{i=1}^m x_i}{2\sigma^2} + \frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow \mu_m = \sigma_m^2 \left(\frac{\sum_{i=1}^m x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \text{ or}$$

$$\mu_m = \sigma_m^2 \left(\frac{m\hat{\mu}_{MLE}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \mu_m =$$

$$\hat{\mu}_{MLE} \times m = \sum_{i=1}^m x_i$$

Detailed derivation (contd.)

Our reference equality:

$$\exp\left(\frac{-1}{2\sigma^2} \sum_{i=1}^m (x_i - \mu)^2 - \frac{(\mu - \mu_0)^2}{2\sigma_0^2}\right) = \exp\left(\frac{-1}{2\sigma_m^2} (\mu - \mu_m)^2\right),$$

Matching coefficients of μ^2 , we get

$$\frac{-\mu^2}{2\sigma_m^2} = \frac{-\mu^2}{2} \left(\frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}\right) \Rightarrow \frac{1}{\sigma_m^2} = \frac{1}{\sigma^2} + \frac{m}{\sigma^2}$$

Matching coefficients of μ , we get

$$\frac{2\mu\mu_m}{2\sigma_m^2} = \mu \left(\frac{2\sum_{i=1}^m x_i}{2\sigma^2} + \frac{2\mu_0}{2\sigma_0^2}\right) \Rightarrow \mu_m = \sigma_m^2 \left(\frac{\sum_{i=1}^m x_i}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \text{ or}$$

$$\mu_m = \sigma_m^2 \left(\frac{m\hat{\mu}_{ML}}{\sigma^2} + \frac{\mu_0}{\sigma_0^2}\right) \Rightarrow \mu_m = \left(\frac{\sigma^2}{m\sigma_0^2 + \sigma^2} \mu_0\right) + \left(\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2} \hat{\mu}_{ML}\right)$$

How abt: $W \in \mathbb{R}^n$
st $w_i \in \mathcal{N}(0, \frac{1}{\lambda})$

$w_t \rightarrow 0$
as $m \rightarrow \infty$

$w_t \rightarrow 1$
as $m \rightarrow \infty$

Data starts dominating

Summary: Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \dots x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\sigma_{MLE}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?
- Answer: $\Pr(\mu | x_1 \dots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that

Summary: Conjugate Prior for (univariate) Gaussian

- Let $\Pr(X) \sim \mathcal{N}(\mu, \sigma^2)$ and let the data $\mathcal{D} = x_1 \dots x_m$
- $\mu_{MLE} = \frac{1}{m} \sum_{i=1}^m x_i$ and $\sigma_{MLE}^2 = \frac{1}{m} \sum_{i=1}^m (x_i - \mu_{MLE})^2$
- Suppose you are told that the conjugate prior for the (univariate) normally distributed random variable X in the case that σ^2 is not a random variable is $\Pr(\mu) = \mathcal{N}(\mu_0, \sigma_0^2)$. Then the **posterior** is?
- Answer: $\Pr(\mu | x_1 \dots x_m) = \mathcal{N}(\mu_m, \sigma_m^2)$ such that
- $\mu_m = \left(\frac{\sigma^2}{m\sigma_0^2 + \sigma^2} \mu_0 \right) + \left(\frac{m\sigma_0^2}{m\sigma_0^2 + \sigma^2} \hat{\mu}_{ML} \right)$
- $\frac{1}{\sigma_m^2} = \frac{1}{\sigma_0^2} + \frac{m}{\sigma^2}$

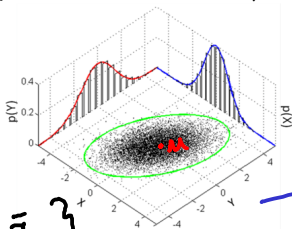
Multivariate Normal Distribution and MLE estimate

- ① The multivariate Gaussian (Normal) Distribution is: $(x \in \mathbb{R}^n)$

$$\mathcal{N}(\mathbf{x}; \underline{\mu}, \underline{\Sigma}) = \frac{1}{(2\pi)^{\frac{n}{2}} |\underline{\Sigma}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\underline{\mu})^T \underline{\Sigma}^{-1}(\mathbf{x}-\underline{\mu})}$$

when $\underline{\Sigma} \in \mathbb{R}^{n \times n}$ is

positive-definite and $\underline{\mu} \in \mathbb{R}^n$



$$P_k(x_1) = \int_{x_2, \dots, x_n} \mathcal{N}(x, \underline{\mu}, \underline{\Sigma}) dx_2 \dots dx_n = \mathcal{N}(x_1, \mu_1, \sigma_1^2)$$

plane of $x \in \mathbb{R}^2$
points scattered around
 $\underline{\mu}$ as per 3- σ rule

$$D = \{\bar{x}_1, \dots, \bar{x}_m\}$$

② $\underline{\mu}_{MLE} = \frac{1}{m} \sum_{i=1}^m \mathbf{x}_i \sim \frac{1}{m} \sum_{i=1}^m \phi(\mathbf{x}_i)$ and

$$\underline{\Sigma}_{MLE} \sim \frac{1}{m} \sum_{i=1}^m (\phi(\mathbf{x}_i) - \underline{\mu}_{MLE})(\phi(\mathbf{x}_i) - \underline{\mu}_{MLE})^T = [:] E [:]$$

$$(\underline{\Sigma}_{MLE})_{kj} = \frac{1}{m} \sum_{i=1}^m (x_{ik} - (\underline{\mu}_{MLE})_k)(x_{ij} - (\underline{\mu}_{MLE})_j)$$

Weka workbench for
ML: visualization in
pairs of (x_i, x_j)

Summary for MAP estimation with Normal Distribution

Univariate Case

- Summary: With $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$

$$\frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{\mu_m}{\sigma_m^2} = \frac{m}{\sigma^2} \hat{\mu}_{mle} + \frac{\mu_0}{\sigma_0^2}$$

such that $p(x|D) \sim \mathcal{N}(\mu_m, \sigma_m^2)$. Here n/σ^2 is due to noise in observation while $1/\sigma_0^2$ is due to uncertainty in μ

- For the Bayesian setting for the multivariate case with fixed Σ

$$x \sim \mathcal{N}(\mu, \Sigma), \mu \sim \mathcal{N}(\mu_0, \Sigma_0) \text{ \& } p(x|D) \sim \mathcal{N}(\mu_m, \Sigma_m)$$

$$\Sigma \in \mathbb{R}^{n \times n} \quad \mu \in \mathbb{R}^n \quad \mu_0 \in \mathbb{R}^n \quad \Sigma_0 \in \mathbb{R}^{n \times n}$$

$$\Sigma_m^{-1} = m \Sigma^{-1} + \Sigma_0^{-1}$$

$$\mu_m \Sigma_m^{-1} = m \hat{\mu}_{mle} \Sigma^{-1} + \mu_0 \Sigma_0^{-1}$$

if $n=1$

$$\Sigma_m = (\sigma_m)^2$$

$$\Sigma_m^{-1} = \left(\frac{1}{\sigma_m}\right)^2$$

Summary for MAP estimation with Normal Distribution

- Summary: With $\mu \sim \mathcal{N}(\mu_0, \sigma^2_0)$ and $x \sim \mathcal{N}(\mu, \sigma^2)$

$$\frac{1}{\sigma_m^2} = \frac{m}{\sigma^2} + \frac{1}{\sigma_0^2}$$

$$\frac{\mu_m}{\sigma_m^2} = \frac{m}{\sigma^2} \hat{\mu}_{mle} + \frac{\mu_0}{\sigma_0^2}$$

such that $p(x|D) \sim \mathcal{N}(\mu_m, \sigma_m^2)$. Here n/σ^2 is due to noise in observation while $1/\sigma_0^2$ is due to uncertainty in μ

- For the Bayesian setting for the multivariate case with fixed Σ
 $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, $\mu \sim \mathcal{N}(\mu_0, \Sigma_0)$ & $p(\mathbf{x}|D) \sim \mathcal{N}(\mu_m, \Sigma_m)$

$$\Sigma_m^{-1} = m\Sigma^{-1} + \Sigma_0^{-1}$$

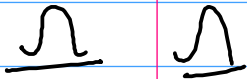
$$\Sigma_m^{-1} \mu_m = m\Sigma^{-1} \hat{\mu}_{mle} + \Sigma_0^{-1} \mu_0$$

- We now conclude our discussion on Bayesian Linear Regression..

$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$

$$[w_1 \dots w_n] = \bar{w} = \mathcal{N}(0, \frac{1}{\lambda} \mathbf{I})$$

$$\rightarrow \Sigma = \begin{bmatrix} \frac{1}{\lambda} & & 0 \\ & \ddots & \\ 0 & & \frac{1}{\lambda} \end{bmatrix}$$



$w_i \perp w_j$ (w_i is independent of w_j)

$$\text{cov}(w_i, w_j) = 0$$

Prior Distribution for \mathbf{w} for Linear Regression

$$y = \mathbf{w}^T \phi(x) + \varepsilon$$
$$\varepsilon \sim \mathcal{N}(0, \sigma^2)$$

- We saw that when we try to maximize log-likelihood we end up with $\hat{\mathbf{w}}_{MLE} = (\phi^T \phi)^{-1} \phi^T \mathbf{y}$
- We can use a Prior distribution on \mathbf{w} to avoid over-fitting

$$w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$$

..Each component w_i is approximately bounded within $\pm \frac{3}{\sqrt{\lambda}}$.
 λ is also called the precision of the Gaussian

- Q1: How do deal with Bayesian Estimation for Gaussian distribution?
- Q2: Then what is the (collective) prior distribution of the n -dimensional vector \mathbf{w} ?

Multivariate Normal Distribution and MAP estimate

Recall: $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi|\boldsymbol{\Sigma}|)^{n/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$

$\boldsymbol{\Sigma}^{-1} = \lambda \mathbf{I}, |\boldsymbol{\Sigma}| = \frac{1}{\lambda^n}$

① If $w_i \sim \mathcal{N}(0, \frac{1}{\lambda})$ then $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \frac{1}{\lambda} \mathbf{I})$ where \mathbf{I} is an $n \times n$ identity matrix

② \Rightarrow That is, \mathbf{w} has a multivariate Gaussian distribution
 $\Pr(\mathbf{w}) = \frac{1}{(\frac{2\pi}{\lambda})^{n/2}} e^{-\frac{\lambda}{2} \|\mathbf{w}\|_2^2}$ with $\boldsymbol{\mu}_0 = \mathbf{0}$. $\boldsymbol{\Sigma}_0 = \frac{1}{\lambda} \mathbf{I}$

③ We will specifically consider Bayesian Estimation for multivariate Gaussian (Normal) Distribution on \mathbf{w} :

$$\frac{1}{(\frac{2\pi}{\lambda})^{n/2}} e^{-\frac{\lambda}{2} \|\mathbf{w}\|_2^2}$$

Substitute for $\boldsymbol{\Sigma}_0 = \frac{1}{\lambda} \mathbf{I}$

$$[\epsilon_1, \dots, \epsilon_m] \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}) \xrightarrow{\quad} \boldsymbol{\Sigma}$$

& determine $\mathcal{P}_1(\mathbf{w}|\mathcal{D})$

} $\mathcal{H}(\mathbf{w})$: