# A NOTE ON BAYESIAN LINEAR REGRESSION [1–3]

*Mohammad Shekaramiz and Todd K. Moon*

Electrical and Computer Engineering Department and Information Dynamics Laboratory
Utah State University
{mohammad.shekaramiz@aggiemail.usu.edu, todd.moon@usu.edu}

**Abstract– In this report, we briefly discuss Bayesian linear regression as well as the proof for the inference to perform prediction based on the training data using this technique.**

## 1. MODEL DESCRIPTION

- **Training data:** input-output pairs $D = \{(\mathbf{x}_m, y_m) | m = 1, 2, ..., M\}$
- Each input is a vector $\mathbf{x}_m$ of dimension $N$
- Suppose that the training data $D$ be a set of *i.i.d.* samples from some unknown distribution
- The standard probabilistic interpretation of linear regression states that

$$y_m = \boldsymbol{\theta}^T \mathbf{x}_m + \epsilon_m, \quad m = 1, 2, ..., M, \tag{1}$$

where $\epsilon$ is the noise and $\epsilon_m \sim \mathcal{N}(0, \sigma^2)$

- For notational convenience, define

$$X := \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_m^T \end{pmatrix}, \mathbf{y} := \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}, \text{ and } \boldsymbol{\epsilon} := \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix},$$

where $X \in \mathbf{R}^{M \times N}$, $\mathbf{y} \in \mathbf{R}^M$, and $\epsilon \in \mathbf{R}^M$.

- In Bayesian linear regression, we assume that a “prior distribution” over parameters is given
- Prior over the weight vector $\boldsymbol{\theta}$:

$$\boldsymbol{\theta} \sim \mathcal{N}(\mathbf{0}, \tau^2 I_N) \tag{2}$$

## 2. POSTERIOR DISTRIBUTION OVER $\theta$

- Apply Bayes’ theorem to obtain the posterior distribution on the weight set $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}|X, \mathbf{y}) \propto p(\mathbf{y}|X, \boldsymbol{\theta}) p(\boldsymbol{\theta}) \tag{3}$$

- Use equations (2) and (3) to find the posterior distribution over $\boldsymbol{\theta}$

$$\log p(\boldsymbol{\theta}|X, \mathbf{y}) \propto$$
$$(\mathbf{y} - X\boldsymbol{\theta})^T (\sigma^2 I_M)^{-1} (\mathbf{y} - X\boldsymbol{\theta}) + \boldsymbol{\theta}^T (\tau^2 I_N)^{-1} \boldsymbol{\theta}$$

- Collect the terms dependent on $\boldsymbol{\theta}$

$$\log p(\boldsymbol{\theta}|X, \mathbf{y}) \propto$$
$$\boldsymbol{\theta}^T (\frac{1}{\sigma^2} X^T X + \frac{1}{\tau^2} I_N) \boldsymbol{\theta} - 2\boldsymbol{\theta}^T (\frac{1}{\sigma^2} X^T \mathbf{y})$$

- Therefore, the posterior distribution over $\boldsymbol{\theta}$ becomes

$$p(\boldsymbol{\theta}|X, \mathbf{y}) = \mathcal{N}(\mu_{\boldsymbol{\theta}}, \Sigma_{\boldsymbol{\theta}}), \tag{4}$$

where

$$\mu_{\boldsymbol{\theta}} = (\frac{1}{\sigma^2} X^T X + \frac{1}{\tau^2} I_N)^{-1} \frac{1}{\sigma^2} X^T \mathbf{y} \tag{5}$$

and

$$\Sigma_{\boldsymbol{\theta}} = (\frac{1}{\sigma^2} X^T X + \frac{1}{\tau^2} I_N)^{-1} \tag{6}$$

- Define $A := \Sigma_{\boldsymbol{\theta}}^{-1}$ to be compatible with the notations used in [1]
- Posterior distribution over $\boldsymbol{\theta}$ becomes

$$p(\boldsymbol{\theta}|X, \mathbf{y}) = \mathcal{N}(\frac{1}{\sigma^2} A^{-1} X^T \mathbf{y}, A^{-1}) \tag{7}$$

## 3. PREDICTION USING BAYESIAN LINEAR REGRESSION

- Assume that there is the same noise model on the testing point, $\{\mathbf{x}_\star, y_\star\}$, as our training points

- Posterior predictive distribution over $y_\star$:
  Integrate out the weight vector $\boldsymbol{\theta}$

$$p(y_\star|\mathbf{x}_\star, X, \mathbf{y}) = \int_{\boldsymbol{\theta}} p(y_\star|\mathbf{x}_\star, \boldsymbol{\theta}) p(\boldsymbol{\theta}|X, \mathbf{y}) d\boldsymbol{\theta}$$
$$= \int_{\boldsymbol{\theta}} \exp\big\{ -[(y_\star - \mathbf{x}_\star^T\boldsymbol{\theta})^T(\sigma^2 I)^{-1}(y_\star - \mathbf{x}_\star^T\boldsymbol{\theta}) +$$
$$(\boldsymbol{\theta} - \frac{1}{\sigma^2}A^{-1}X^T\mathbf{y})^T A(\boldsymbol{\theta} - \frac{1}{\sigma^2}A^{-1}X^T\mathbf{y})]\big\} d\boldsymbol{\theta} \tag{8}$$

- Matrix $A$ is symmetric (covariance matrix), hence

$$p(y_\star|\mathbf{x}_\star, X, \mathbf{y}) =$$
$$\int_{\boldsymbol{\theta}} \exp\big\{ -[\boldsymbol{\theta}^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)\boldsymbol{\theta} - 2\boldsymbol{\theta}^T\frac{1}{\sigma^2}(\mathbf{x}_\star y_* + X^T\mathbf{y})$$
$$+ (\frac{1}{\sigma^2}y_\star^T y_\star + \frac{1}{\sigma^4}\mathbf{y}^T X A^{-1} X^T \mathbf{y})]\big\} d\boldsymbol{\theta}$$

- Results in

$$P(y_\star|\mathbf{x}_\star, X, \mathbf{y}) =$$
$$e^{-[\frac{1}{\sigma^2}(y_\star^T y_\star + \frac{1}{\sigma^2}\mathbf{y}^T X A^{-1} X^T \mathbf{y}) - \frac{1}{\sigma^4}(\mathbf{x}_\star y_\star + X^T\mathbf{y})^T (A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}(-)]}$$
$$\times \int_{\boldsymbol{\theta}} e^{-[(\boldsymbol{\theta} - (A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}\frac{1}{\sigma^2}(\mathbf{x}_\star y_\star + X^T\mathbf{y}))^T (A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)(-)]} d\boldsymbol{\theta}$$

- Collect terms that only depend on $y_\star$

$$\log p(y_\star|\mathbf{x}_\star, X, \mathbf{y}) \propto$$
$$y_\star^T\big(\frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}\mathbf{x}_\star\big)y_\star$$
$$- 2y_\star^T\big(\frac{1}{\sigma^4}\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}X^T\mathbf{y}\big)$$

- Therefore, the posterior of $y_\star$ is Gaussian with the following terms

$$\Sigma_{y_\star} = \big(\frac{1}{\sigma^2} - \frac{1}{\sigma^4}\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}\mathbf{x}_\star\big)^{-1} \tag{9}$$

  and

$$\mu_{y_\star} = \Sigma_{y_\star}\frac{1}{\sigma^4}\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}X^T\mathbf{y}. \tag{10}$$

- Remark: Matrix Inversion Lemma for equation

$$\begin{aligned}&(A + BCD)^{-1} = \\ &A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}\end{aligned} \tag{11}$$

- Simplifying covariance matrix $\Sigma_{y_\star}$ in (9) using (11)

$$\Sigma_{y_\star} = \big((\sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star)^{-1}\big)^{-1}$$

$$\Sigma_{y_\star} = \sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star \tag{12}$$

- Simplifying the mean $\mu_{y_\star}$ in (10) using (11):

$$\begin{aligned}\mu_{y_\star} =& \Sigma_{y_\star}\frac{1}{\sigma^4}\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}X^T\mathbf{y} \\ =& \frac{1}{\sigma^4}(\sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star)\mathbf{x}_\star^T(A + \frac{1}{\sigma^2}\mathbf{x}_\star\mathbf{x}_\star^T)^{-1}X^T\mathbf{y} \\ =& \frac{1}{\sigma^4}(\sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star)\mathbf{x}_\star^T(A^{-1} - \frac{A^{-1}\mathbf{x}_\star\mathbf{x}_\star^T A^{-1}}{\sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star})X^T\mathbf{y} \\ =& \frac{1}{\sigma^4}(\sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star)\mathbf{x}_\star^T A^{-1} - \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star\mathbf{x}_\star^T A^{-1})X^T\mathbf{y}\end{aligned}$$

- Therefore,

$$\mu_{y_\star} = \frac{1}{\sigma^2}\mathbf{x}_\star^T A^{-1}X^T\mathbf{y} \tag{13}$$

## 4. SUMMARY ON THE PREDICTION

$$p(y_\star|\mathbf{x}_\star, X, \mathbf{y}) = \mathcal{N}(\mu_{y_\star}, \Sigma_{y_\star}),$$

where

$$\mu_{y_\star} = \frac{1}{\sigma^2}\mathbf{x}_\star^T A^{-1}X^T\mathbf{y}$$

and

$$\Sigma_{y_\star} = \sigma^2 + \mathbf{x}_\star^T A^{-1}\mathbf{x}_\star.$$

## 5. REFERENCES


[1] C. Rasmussen and C. Williams, *Gaussian Processes in Machine Learning*. MIT Press, 2006.

[2] C. B. Do, "Gaussian processes," Dec. 2007. [Online]. Available: http://see.stanford.edu/materials/aimlcs229/cs229-gp.pdf/

[3] M. Shekaramiz, *Sparse Signal Recovery Based on Compressive Sensing and Exploration Using Multiple Mobile Sensors*. PhD Dissertation, Utah State University, Digitalcommons, 2018.