Small Problem 1: Bayesian Linear Regression

The file “problem-1-generator.R” contains R code to generate the true regression coefficients and the input training data. The model is

$$\mathbf{\Sigma}_1 = 2\boldsymbol{I}_{5\times 5}$$

$$\boldsymbol{\mu} \sim \mathrm{Norm}\big((0,0,0,0,0), \mathbf{\Sigma}_1\big)$$

$$\mathbf{\Sigma}_2 = \boldsymbol{I}_{5\times 5}$$

$$\mathbf{\Sigma}_{prior} \sim \mathrm{Wishart}(1, \mathbf{\Sigma}_2)$$

$$\boldsymbol{w} \sim \mathrm{Norm}\big(\boldsymbol{\mu}, \mathbf{\Sigma}_{prior}^{-1}\big)$$

$$\boldsymbol{x}_{ij} \sim \mathrm{Uniform}(-1, +1) \text{for } j = 1, \ldots, 5; i = 1, \ldots, 500$$

$$\tau \sim \text{ gamma}(0.5, 2)$$

$$\epsilon_i \sim Norm\left(0, \frac{1}{\tau}\right) \text{ for } i = 1, \ldots, 500$$

$$y_i = \boldsymbol{w}^\top \boldsymbol{x}_{ij} + \epsilon_i \text{ for } i = 1, \ldots, 500$$

The file contains 500 training examples generated from a single run of the R code. There are four covariates generated uniformly from $[-1, +1]$. The values of the variables that generated the data are

$$\boldsymbol{\mu} = (-1.8195312, 1.2237587, 0.8361809, -2.6017006, -2.3574193)$$

$$\mathbf{\Sigma}_{prior} = \text{see “problem-1-prior.Sigma.csv”}$$

$$\boldsymbol{w} = (-1.731855, 2.986017, 2.698284, -3.591651, -3.714157)$$

$$(\boldsymbol{x}_i, y_i) \text{ for } i = 1, \ldots, 500 \text{ see “problem-1-data.csv”}$$

Queries/Metrics:

1. Let $P(\widehat{\boldsymbol{w}}|D)$ be the posterior distribution of the estimated weight vector. One metric is the expected squared error $\mathbb{E}[\|\widehat{\boldsymbol{w}} - \boldsymbol{w}\|^2]$ under this distribution.
2. We have provided samples generated from the true posterior distribution $P_{true}(\widehat{\boldsymbol{w}}|D)$. We can estimate the total variation distance between the true distribution and your estimate $P(\widehat{\boldsymbol{w}}|D)$ using the samples generated by your estimated distribution:
$$\int_{\boldsymbol{w}} |P(\widehat{\boldsymbol{w}}|D) - P_{true}(\widehat{\boldsymbol{w}}|D)| d\boldsymbol{w}$$