GEORGE MASON UNIVERSITY
Systems Engineering and Operations Research
OR750/610: Deep Learning, Fall Semester 2019: Homework Assignment 3. Due: XXX (before class)

1. Gibbs Sampler Suppose we model data using the following model

$$\begin{aligned} y_i \sim & N(\mu, \tau^{-1}) \\ \mu \sim & N(0,1) \\ \tau \sim & N(2,1). \end{aligned}$$

The goal is to implement a Gibbs sample for the posterior $\mu, \tau | y$, where $y = (y_1, \ldots, y_n)$ is the observed data. Gibbs sampler algorithms iterates between two steps

1. Sample μ_i from $\mu \mid \tau_{i-1}, y$
2. Sample τ_i from $\tau_i \mid \mu_i, y$

Show that those full conditional distributions are given by

$$\begin{aligned} \mu \mid \tau, y \sim & N\left(\frac{\tau n \bar{y}}{1+n\tau}, \frac{1}{1+n\tau}\right) \\ \tau \mid \mu, y \sim & \text{Gamma}\left(2+\frac{n}{2}, 1+\frac{1}{2}\sum_{i=1}^{n}(y_i-\mu)^2\right) \end{aligned}$$

Use formulas for full conditional distributions and implement the Gibbs sampler. The data y is in the file `MCMCexampleData.txt`.

Plot samples from the joint distribution over (μ, τ) on a scatter plot. Plot histograms for marginal μ and τ (marginal distributions).

2. Gibbs Sampler for Lasso Consider a linear regression model

$$y = \mu + x^T \beta + e, \quad e \sim N(0, \sigma^2).$$

Given data $D = (X, y)$, where X is the $n \times p$ matrix of standardized regressors and y is the n-vector of outputs. Implement a Gibbs sampler for this model when Laplace prior is used for model coefficients β_i. Use scale mixture normal representation.

$$\begin{aligned} \beta \mid \sigma^2, \tau_1, \ldots, \tau_p \sim & N(0, \sigma^2 D_\tau) \\ D_\tau = & \text{diag}(\tau_1^2, \ldots, \tau_p^2) \\ \tau_i^2 \mid \lambda \sim & \exp(\lambda^2/2) \\ \sigma^2 \sim & 1/\sigma^2. \end{aligned}$$

It can be shown that `Park08BayesLasso.pdf`

$$\begin{aligned} \beta \mid D, D_\tau \sim & N(A^{-1}X^T y, \sigma^2 A^{-1}), \quad A = X^T X + D_\tau^{-1} \\ \sigma^2 \mid \beta, D, D_\tau \sim & \text{InverseGamma}\left((n-1)/2 + p/2, (y-X\beta)^T(y-X\beta)/2 + \beta^T D_\tau^{-1}\beta/2\right) \\ 1/\tau_j^2 \mid \beta_j, \lambda \sim & \text{InverseGaussian}\left(\sqrt{\frac{\lambda^2\sigma^2}{\beta_j^2}}, \lambda^2\right) \end{aligned}$$

The formulas above assume that X is standartized, e.g. observations for each feature are scaled to be of mean 0 and sd 1 $x^j = (x^j - \bar{x}^j)/sd(x^j)$, where x^j is the jth column of matrix X, and y is centered $y = y - \bar{y}$. Also, note that Python's `scipy.stats.invgauss` samples from InverseGaussian with shape

parameter 1, to scale it, you can use the following property, if $X \sim \text{InverseGaussian}(\mu, \lambda)$, then $tX \sim \text{InverseGaussian}(t\mu, t\lambda)$.

You can use empirical priors and initialize the parameters as follows

$$\begin{aligned}
\beta &= (X^T X + I)^{-1} X^T y \\
r &= y - X\beta \\
\sigma^2 &= r^T r / n \\
\tau^{-2} &= 1/(\beta \odot \beta) \\
\lambda &= p\sqrt{\sigma^2} / \sum |\beta|,
\end{aligned}$$

here n is number of rows (observations) and p is number of columns (inputs) in matrix X.

Example of implementation in R: `https://cs.gmu.edu/~pwang7/code/gibbsBLasso.R`.

Use Gibbs sampler to fit your model to the `longley` data. Data set description is here: `https://stat.ethz.ch/R-manual/R-devel/library/datasets/html/longley.html`. How do your results compare to classical ridge? Include histograms of the posterior distributions of coefficients with the prior density overlaid, plus means and credible intervals. How sensitive are the results to the prior assumptions? How do the estimates of τ_i compare to the best estimate from Ridge?