Due March 13. Show all work and code for full credit. Accepted one day late for half credit.

**Problem 1.** Consider the sparse regression model discussed in class:

\[ y \sim N(Xw, \lambda^{-1}I), \quad w \sim N(0, \text{diag}(\alpha_1, \ldots, \alpha_{d+1})^{-1}), \quad \alpha_k \sim \text{Gam}(a_0, b_0), \quad \lambda \sim \text{Gam}(e_0, f_0). \]

The data is a vector of observations \( y \in \mathbb{R}^N \) and an \( N \times d + 1 \) matrix of covariates \( X \), where the first column of \( X \) is a vector of ones.

**Part (a)** Derive a variational inference algorithm for the variables \( w, \alpha_1, \ldots, \alpha_{d+1} \) and \( \lambda \). Use the factorization

\[ q(w, \alpha_1, \ldots, \alpha_{d+1}, \lambda) = q(w)q(\lambda) \prod_{k=1}^{d+1} q(\alpha_k) \]

for the variational distribution and let each \( q \) be in the same family as the prior. Use the variational lower bound to do this.

**Part (b)** Show that this is the optimal \( q \) distribution given our selected factorization.

**Part (c)** Implement the algorithm in Part (a) and run it on the data provided. Set the prior parameters \( a_0 = b_0 = 10^{-16} \) and \( e_0 = f_0 = 1 \). For each data set:

1. Run your algorithm for 1000 iterations and plot the variational lower bound.

2. Using \( q(w) \), plot the mean and standard deviation of \( p(y|X, \lambda) = \int p(y|X, w, \lambda)q(w)dw \).

   Set \( \lambda = 10^{10} \) for this part in order to focus on the error due to the uncertainty of \( w \). Show the standard deviation as error bars. In the same figure show a scatter plot of the original data and plot the ground truth (see below).

3. Show a stem plot of \( 1/\mathbb{E}_q[\alpha_k] \) as a function of \( k \). Also indicate \( 1/\mathbb{E}_q[\lambda] \) somewhere.

Details about the data:

The data was generated by sampling \( z \sim \text{Uniform}(-5, 5) \) independently \( N \) times for \( N = 100, 250, 500 \) (giving a total of three data sets). For each \( z_n \) in a given data set, the response \( y_n = 10 \ast \text{sinc}(z_n) + \epsilon_n \), where \( \epsilon_n \sim N(0, 1) \). The ground truth mentioned above is \( 10 \ast \text{sinc}(\cdot) \).

We use \( z_n \) to construct the “kernel matrix” \( X \). This is a mapping of \( z_n \) into a higher dimensional space (see Bishop for more details). For our purposes, it’s just important to know that the \( n \)th row of \( X \) corresponds to the location \( z_n \). We let \( X_{n,1} = 1 \) and use the Gaussian kernel for the remaining dimensions, \( X_{n,i+1} = \exp\{- (z_n - z_i)^2 \} \) for \( i = 1, \ldots, N \). Therefore, the dimensionality is one greater than the number of data points. The sparse model picks out the relevant locations within the data for performing the regression.

Each data set contains the vector \( y \), the matrix \( X \) and the vector of original locations \( z \). This last vector will be useful for plotting.