Problem 1. (Monte Carlo integration / Variational inference) When encountering problems in variational inference, we measure distance between two distributions, $q(x), p(x)$ with the Kullback–Leibler divergence, which is defined as:

$$KL(q||p) = \int \log q(\boldsymbol{x})\, q(x)\, d\boldsymbol{x} - \int \log p(\boldsymbol{x})\, q(\boldsymbol{x})\, d\boldsymbol{x} \tag{1}$$

$$= E_q\{\log(q(\boldsymbol{X})) - \log p(\boldsymbol{X})\}$$

For complex distributions, this integral is not easily accessible. We will now consider a target distribution being the bivariate normal distribution with unit variance and mean zero. This distribution is defined as:

$$p(\boldsymbol{x}) = \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}\boldsymbol{x}^{\boldsymbol{T}}\Sigma^{-1}\,\boldsymbol{x}\right\} \tag{2}$$

where

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}. \tag{3}$$

In a standard approach to variational inference we consider the mean field approximation where we look for a solution of the form:

$$q(x) = q_1(x_1) \cdot q_2(x_2). \tag{4}$$

Which means that we are looking for a distribution with independent components, specifically we are looking for solutions where $q_j(x_j) = \phi(x_j; \mu_j, \sigma_j^2)$, with $\phi(x; a, b^2)$ being the normal density; mean a and variance b^2. Hint: You can use a library to evaluate the multivariate density, e.g. `dmvnorm` in the `mvtnorm` package.

a) Let $\rho = 0.9$, $\mu_1 = \mu_2 = 0$ and $\sigma_1 = \sigma_2 = 1$.
   Make a function which evaluates the Kullback–Leibler divergence $KL(q||p)$ using Monte Carlo integration. How many samples do you need to have a result with Monte Carlo variability less than 0.01? (Here variability is measured in terms of standard deviation.)

b) It is possible to show that the Kullback–Leibler divergence has its minimum for $\mu_1 = \mu_2 = 0$, and $\sigma_1 = \sigma_2 = \sqrt{1-\rho^2}$. In light of this result, discuss strengths and weaknesses of an analysis based on mean field variational inference.

The rest of problem 1 is for STK 9051 only. An algorithm for deriving the mean field approximation in the general case, is to use coordinate ascent variational inference, i.e. the CAVI algorithm. A version of the CAVI algorithm is detailed below.

Algorithm 1 (CAVI)

1) Initialize $q_2^{(0)}(x) = \phi(x; a, b^2)$,
2) While not converged iterate:
   a. Set $q_1^{(i)}(x_1) \propto \exp\{E_{q_2^{(i-1)}}(\log p(x_1|x_2))\}$
   b. Set $q_2^{(i)}(x_2) \propto \exp\{E_{q_1^{(i)}}(\log p(x_2|x_1))\}$

c) (STK 9051 only). We will now analyze use of this algorithm for the distribution $p(\boldsymbol{x})$ in (2). Start off with $q_2^{(0)}(x) = \phi(x; 1, 1^2)$. Compute $q_1^{(1)}(x)$, and derive $q_j^{(n)}(x)$. Comment on the result. You can use without proof (if needed) that:

$$\log p(x_i|x_j) = const\ -\frac{1}{2}\frac{(x_i - \rho x_j)^2}{1-\rho^2} \tag{5}$$

Problem 2 (Model selection). The problem of model selection is NP- hard, thus we need heuristic algorithms to find good solutions. The genetic algorithm is inspired by natural selection.

a) Describe the main elements in a Genetic algorithm.

b) Implement a genetic algorithm for identifying the optimal model using the AIC "An information Criterion" to select model. You can use the file `baseball_genetic.R` on the course webpage for inspiration, but retain only the parts the code that you actually present in text. That is explain each part in the algorithm, such that it can be recognized in the code. Apply the code to the diabetes data set `diabetesEx2.dat`. [no extra points for using a complex solution]

c) The problem of model selection, is standard in statistics. When considering Bayesian model selection, the standard McMC methods can't be applied directly. Which feature of model

selection is it that complicates the situation for standard McMC algorithms? Do you know any algorithms which can be used to solve this problem?

Problem 3, (EM algorithm) We consider iid data, $y_i, i = 1, \dots, n,$ from a mixture distribution

$$p(y) = \nu \cdot \phi(y; \mu, 1) + (1 - \nu)\phi(y; -\mu, 1) \quad (6)$$

Where $0 < \nu < 1$, $\phi(x; a, b^2)$ denote the normal density, with mean a and variance b^2, and the parameter μ gives the separation of the modes. We will now derive the estimate for the parameters $\theta = (\nu, \mu)$, using the EM algorithm.

a) Set up the expression for the complete likelihood using a hidden variable C_i, which indicates the class membership of data y_i.

b) Give an expression for the $Q(\theta|\theta^{(t)})$ function, and use this to set up the EM algorithm for estimation. Show in particular that :

$$P\left(C_i \middle| y_i, \nu^{(t)}, \mu^{(t)}\right) = \frac{\nu^{(t)}\phi(y_i, \mu^{(t)}, 1)}{\nu^{(t)}\phi(y_i, \mu^{(t)}, 1) + (1 - \nu^{(t)})\phi(y_i, -\mu^{(t)}, 1)} \quad (7)$$

c) Implement the estimator and apply it to the dataset `EM_mixture.dat`. What are the final estimates?

d) In the case where the class membership is known the standard deviation of the estimators are 0.016 and 0.040 for ν and μ respectively. The estimators are also independent. Implement the bootstrap estimate of the uncertainty of the parameters using the EM algorithm. Make a scatterplot of 500 resamples. Comment on the results, also in relation to the uncertainty in the complete likelihood.

Problem 4(Gibbs sampler) Problem 3's parameter estimation, can also be solved using Bayesian methodology. The posterior distribution can be sampled using a Gibb sampler with a set of augmented variables. Introduce the class memberships as an augmented variable set $C_i, i = 1, \dots, n.$

a) Derive the expressions for $P(C_i|y_i, \nu, \mu)$, $p(\nu|y_i, C_i, \mu)$ and, $p(\mu|y_i, C_i, \nu)$. Show how this can be used to sample the posterior distribution of ν and μ. [Hint: use results from problem 3, and use properties of conditional independence when conditioning to membership

variables $C_i, i = 1, \dots, n$. The probability density of the beta distribution $f(x; a, b)$ with shape parameters (a, b) has the form: $f(x; a, b) \propto x^{a-1}(1-x)^{b-1}$.]

b) Implement the Gibbs sampler and apply it to the data set `EM_mixture.dat`. Select the prior distribution as you see fit, but give an argument for your choice. Let the number of samples of each variable be 1100.

c) Discuss the two concepts “burn in” and “effective sample size". Is a bur in of 100 sample sufficient for your chain above? Compute the effective sample sizes for μ and ν.

Problem 5 (STAN) The library rstan uses Hamiltonian Monte Carlo as a generic tool for implementing Bayesian inference. The stan program below defines a statistical model with data x, and parameters nu, mu1, and mu2. The stan program is also available in the file `Oppg5.stan`.

```
data{
int<lower=1> N;
real X[N];
}

parameters{
real<lower=0,upper=1> nu;
real <lower=0> mu1;
real <upper=0> mu2;
}

model{
for(i in 1:N){
  target+=log(nu*exp(normal_lpdf(X[i]|mu1,1))+(1-nu)*exp(normal_lpdf(X[i]|mu2,1)));

}
}
```

a) The program implies a prior for the parameters and a likelihood for the data. State these statistical models.

b) Run the model using N=625, and X from `EM_mixture.dat`. Show a scatterplot of mu1 and mu2.