Solutions to problems: STK 4051 / STK 9051 exam spring 2021

Problem 1

a) The principle of probability transform. If $F(x) = P(X \leq x)$ is the cumulative distribution of X, and $u_1, u_2, \ldots, u_n$ are samples from a uniform distribution uniform on the unit interval, then the set of samples defined by $x_i = F^{-1}(u_i)$ are samples from the distribution of X. Formalized:

$$P(F^{-1}(U) < x) = P\big(U \leq F(x)\big) = F(x)$$

Where we have used, properties of a monotone transform, and the property of the cumulative distribution for a uniform variable. Here we have:

$$u = F(x) = 1 - \left(\frac{x_m}{x}\right)^{\alpha}, x > x_m$$

$$\frac{1}{1-u} = \left(\frac{x}{x_m}\right)^{\alpha}$$

$$x = F^{-1}(u) = \frac{x_m}{(1-u)^{\frac{1}{\alpha}}}$$

Note that since $1 - U$ also is uniform then we can also write:

$$x = x_m \cdot u^{-\frac{1}{\alpha}}$$

b) In rejection sampling we need a bound on the ratio $f(x)/g(x)$. The function has three regions left central and right. The left and right are symmetric, so it suffices to discuss only one side and the central part.

Central region: $|x| < 1$

$$\max_{-1 \leq x \leq 1} \frac{f(x)}{g(x)} = \max_{-1 \leq x \leq 1} \frac{4}{\pi(1+x^2)} = \frac{4}{\pi}$$

Right side: $x > 1$

$$R(x, \alpha) = \frac{f(x)}{g(x)} = \frac{4x^{\alpha+1}}{\alpha \cdot \pi\,(1+x^2)} = \left(\frac{4}{\alpha \cdot \pi}\right)\frac{x^{\alpha-1}}{1+\frac{1}{x^2}}$$

Case $\alpha > 1$: limit $x \to \infty \implies R(x) \to \infty$, The ratio is unbounded, thus the distribution can not be used for rejection sampling.

Case $\alpha = 1$: $\max_{x} \left(\frac{4}{\pi}\right) \cdot \frac{1}{1+\frac{1}{x^2}} \leq 4/\pi$ The function is monotone increasing towards the bound.

Case $0 < \alpha < 1$: There are many ways to solve this. The bound do not need to be tight. One way is: $x^{\alpha-1}/(1+\frac{1}{x^2}) < 1$, for $x > 1$ Thus

$$e(x) = \frac{4}{\alpha \cdot \pi} g(x) \text{ if } 0 < \alpha \leq 1$$

The acceptance rate is the inverse of the bound: $\frac{\pi\alpha}{4}$

A tight bound is found by: $\frac{\partial R(x,\alpha)}{\partial x} = \frac{x^{\alpha}(\alpha+1+(\alpha-1)x^2)}{(1+x^2)^2}$, the derivative is positive until x^2 becomes large enough then it remains negative thus the zero crossing is a maximum, zero crossing for $x_\alpha = \sqrt{\frac{\alpha+1}{1-\alpha}}$. The maximum value is $R(x_\alpha, \alpha)$. The envelope becomes:

$$e(x) = \begin{cases} \frac{4}{\pi} g(x) & \text{if } \alpha = 1 \\ \max\left\{\frac{4}{\pi}, R(x_\alpha, \alpha)\right\} g(x) & \text{if } 0 < \alpha < 1 \end{cases}$$

The acceptance rate is the inverse bound.

$$\text{acceptance rate} = \begin{cases} \frac{\pi}{4} & \text{if } \alpha = 1 \\ \left(\max\left\{\frac{4}{\pi}, R(x_\alpha, \alpha)\right\}\right)^{-1} & \text{if } 0 < \alpha < 1 \end{cases}$$

Problem 2

The complete log likelihood

$$l(\theta|\boldsymbol{x}, \boldsymbol{c}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \log\frac{1}{K} I(c_i = k) \log\phi(x_i; \mu_k, \sigma_k^2)$$
$$= \sum_{i=1}^{n}\sum_{k=1}^{K} \log\frac{1}{K} I(c_i = k)\left(-\frac{1}{2}\log(2\pi\sigma_k^2) - \frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}\right)$$

$$Q(\theta|\theta^{(t)}) = E\left(l(\theta|\boldsymbol{x}, \boldsymbol{c})|\boldsymbol{x}, \theta^{(t)}\right)$$
$$= \text{Const} + \sum_{i=1}^{n}\sum_{k=1}^{K} P(C_i = k|\boldsymbol{x}, \theta^{(t)})\left(-\frac{1}{2}\log(\sigma_k^2) - \frac{1}{2}\frac{(x_i - \mu_k)^2}{\sigma_k^2}\right)$$

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial \mu_k} = \sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})\left(\frac{(x_i - \mu_k)}{\sigma_k^2}\right) = 0$$

$$\mu_k \sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)}) = \sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)}) x_i$$

$$\mu_k = \frac{\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)}) x_i}{\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})}$$

$$\frac{\partial Q(\theta|\theta^{(t)})}{\partial \sigma_k^2} = \sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})\left(-\frac{1}{2\sigma_k^2} + \frac{1}{2}\frac{(x_i - \mu_k)^2}{(\sigma_k^2)^2}\right) = 0$$

$$\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)}) = \frac{1}{\sigma_k^2}\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})(x_i - \mu_k)^2$$

$$\sigma_k^2 = \frac{\sum_{i=1}^{n} P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)(x_i - \mu_k)^2}{\sum_{i=1}^{n} P(C_i = k|\boldsymbol{x}, \theta^{(t)})}$$

$$P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right) = \frac{p(C_i = k, X_i = x_i|\, \theta^{(t)})}{p(X_i = x_i|\theta^{(t)})} = \frac{\frac{1}{K}\phi\left(x_i; \mu_k, \sigma_k^2\right)}{\frac{1}{K}\sum_{m=1}^{K}\phi(x_i; \mu_m, \sigma_m^2)} = \frac{\phi\left(x_i; \mu_k, \sigma_k^2\right)}{\sum_{m=1}^{K}\phi(x_i; \mu_m, \sigma_m^2)}$$

b) The change is that the probability weigth in the EM estimator becomes one if the class match the information, zero otherwise. Thus for data with information of class we replace $P\left(C_i = k|\boldsymbol{x}, \theta^{(t)}\right)$ with $I(C_i = c_i)$.

In terms of the expression for Q, we have that the expectation is taken over the observed data. Thus when the class is observed, we get perfect information about the class, which gives the indicator function in the sum.

c) Nonparametric bootstrap: Generate new data sets by resampling with replacement the data records from the original set. Positive: The sample is data driven, no assumption of the distributions are used. Negative: The number of known label classes will vary between samples.

Parametric bootstrap: Generate new data by random samples from the estimated model, apply the same label selection strategy as in the main set. Positive: We recreate the mechanism for selecting labels. Negative: We are limited to the model we have fitted.

d) The inverse of the information matrix is an estimator of the sample covariance. Thus it is desirable to have large entries in the information matrix and small values in the inverse.

Strategy A gets some information, it is better than the strategy B, but still close to the unsupervised approach

Strategy B is the worst choice. When we select data from the edges this is where we already are quite certain about the classes, thus this brings little information compared with the unsupervised.

Strategy C gets the most information. This is natural since this gets labels from the region with large overlap in the distribution. We get information closer to the complete information.

The information matrices are sorted in terms of information content

Unsupervised < Strategy B < Strategy A < Strategy C < Complete

Problem 3

a)The target distribution is

$$\pi(x) = p(x|y) = \frac{p(x)p(y|x)}{p(y)}$$

To prove we compute:

$$E\big(Wh(X)\big) = E\left(\frac{p(X)p(y|X)}{g(X)}h(X)\right) = \int_{-\infty}^{\infty}\frac{p(x)p(y|x)}{g(x)}h(x)g(x)dx$$

$$= p(y) \int_{-\infty}^{\infty} \frac{p(x)p(y|x)}{p(y)} h(x) \frac{g(x)}{g(x)}\, dx$$

$$= p(y) \int_{-\infty}^{\infty} \pi(x)h(x)\, dx = p(y)E_{\pi}(h(X))$$

Thus $c = p(y)$.

The estimate of the integral is:

$$\frac{\sum_{i=1}^{M} w_i h(x_i)}{\sum_{i=1}^{M} w_i}$$

b)A sample which is properly weighted with respect to a joint distribution is properly weighted with respect to any of the marginal distributions, since this is just a restriction of the set of functions in the definition. Thus we can compute the weights of $p(x_1, x_2, \dots, x_i | y_1, y_2, \dots, y_i) = p(\boldsymbol{x}|\boldsymbol{y})$. Using the same argument as in a, we find that $c = p(\boldsymbol{y})$.

$$W_i = \frac{p(\boldsymbol{x})p(\boldsymbol{y}|\boldsymbol{x})}{g(\boldsymbol{x})} = \frac{p(x_1)p(y_1|x_1)\prod_{n=2}^{i} p(x_n|x_{n-1})p(y_n|x_n)}{g(x_1)\prod_{n=2}^{i} g(x_n|x_{n-1})}$$

$$= \frac{p(x_1)p(y_1|x_1)}{g(x_1)} \prod_{n=2}^{i} \frac{p(x_n|x_{n-1})p(y_n|x_n)}{g(x_n|x_{n-1})}$$

$$= W_{i-1}\ \frac{p(x_i|x_{i-1})p(y_i|x_i)}{g(x_i|x_{i-1})}$$

Problem 4

a) Burn in up to about 100 samples. After this we are in the target zone. The sample path consists of regions where the parameter is unchanged, this correspond to rejection of the proposal, and thus we expect low acceptance rate. The chain must be irreducible [can visit any place I sample space with a finite number of steps], aperiodic [the sample path do not have a deterministic cycle] and recurrent [we will always return to a set of non-negligible size]. We need to match the

b) The effective number of samples is $N_{\text{eff}} = \frac{N}{1+2\sum_{h\geq 1}\rho(h)}$, so the higher the number in the table is the lower is the effective sample size. In the table there are two trends.

1)The acceptance rate decrease as the magnitude increase.

2) The effective number of samples start off low and increase until about 0.1, 0.2, but for 0.5 it decrease again.

This is a standard behavior of the random walk algorithm.

With a small magnitude of the change, there are many accepts, but each permutation is small, thus many samples are needed to make a significant change in the parameter this results in high correlation in sample and a low number of effective samples.

With a larger magnitude on the change there are more rejects, but the change is larger, thus the autocorrelation decreases resulting in a higher number of effective samples.

When the magnitude increases even more, the number of rejects still increase, but the changes made at each step does not become sufficiently large to balance this, thus there are long periods of repeats in the data set. The situation is as in figure 2. The intervals with repeat create high autocorrelation in the dataset and a lower number of effective sample size.

The final selection should thus be 0.2 which has overall the largest effective number of samples, (or 0.1 which is better for some)

c) Algorithm:
1) Initiate μ η_X and η_R
2) Repeat for a fixed number of samples
   a) Make proposal as detailed above
   b)Accept proposal if $U <$ M − H Ratio, otherwise keep current value. $U \sim$ Uniform[0,1]

In the random walk we have the proposals:

$$\eta_R^{\text{prop}} = \eta_R^{\text{current}} + \epsilon_R, \quad \epsilon_R \sim N(0, \tau^2_{\text{change,R}})$$

$$\eta_X^{\text{prop}} = \eta_X^{\text{current}} + \epsilon_X, \quad \epsilon_X \sim N(0, \tau^2_{\text{change,X}})$$

$$\mu^{\text{prop}} = \mu^{\text{current}} + \epsilon_\mu, \quad \epsilon_\mu \sim N(0, \tau^2_{\text{change,}\mu})$$

We have $\frac{\partial \eta_R}{\partial \sigma_R^2} = \frac{1}{\sigma_R^2} \propto p(\sigma_R^2)$. $\frac{\partial \eta_X}{\partial \sigma_X^2} = \frac{1}{\sigma_X^2} \propto p(\sigma_X^2)$. The acceptance rate becomes:

$$\text{M} - \text{H Ratio} = \frac{L(\mu, \exp \eta_X^{\text{prop}}, \exp \eta_R^{\text{prop}}; \boldsymbol{y})}{L(\mu, \exp \eta_X^{\text{current}}, \exp \eta_R^{\text{current}}; \boldsymbol{y})}$$

The prior is canceled by the Jacobi of the transform, the reverse transition is canceled by the symmetry of the proposal distribution.

d) STK 9051: The joint distribution becomes:

$$p(\sigma_R^2, \sigma_X^2, \mu, \boldsymbol{x}, \boldsymbol{y}) = p(\sigma_X^2) \cdot p(\sigma_X^2) p(\mu) \prod_{i=1}^{n} \phi(x_i, \mu, \sigma_X^2) \prod_{j=1}^{2} \phi(y_{ij}; x_i, \sigma_R^2)$$

$$p(\sigma_R^2 | \sigma_X^2, \mu, \boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{\sigma_R^2} \prod_{i=1}^{n} \prod_{j=1}^{2} \phi(y_{ij}; x_i, \sigma_R^2) = \frac{1}{\sigma_R^2} \prod_{i=1}^{n} \prod_{j=1}^{2} \frac{1}{\sqrt{2\pi\sigma_R^2}} \exp - \frac{1}{2} \left(\frac{(y_{ij} - x_i)^2}{\sigma_R^2} \right)$$

$$\propto \frac{1}{(\sigma_R^2)^{1+n}} \exp - \frac{1}{2\sigma_R^2} \left(\sum_{i=1}^{n} \sum_{j=1}^{2} (y_{ij} - x_i)^2 \right)$$

Is an inverse gamma distribution with parameters: $\alpha = n, \beta = \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{2} (y_{ij} - x_i)^2$

$$p(\sigma_X^2 | \sigma_R^2, \mu, \boldsymbol{x}, \boldsymbol{y}) \propto \frac{1}{\sigma_X^2} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma_X^2}} \exp - \frac{1}{2} \left(\frac{(x_i - \mu)^2}{\sigma_X^2} \right) \propto \frac{1}{(\sigma_X^2)^{1+\frac{n}{2}}} \exp - \frac{1}{2\sigma_X^2} \sum_{i=1}^{n} (x_i - \mu)^2$$

Is an inverse gamma distribution with parameters: $\alpha = \frac{n}{2}, \beta = \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$

$$p(\mu|\sigma_R^2,\sigma_X^2,\boldsymbol{x},\boldsymbol{y}) \propto \prod_{i=1}^{n}\frac{1}{\sqrt{2\pi\sigma_X^2}}\exp-\frac{1}{2}\left(\frac{(x_i-\mu)^2}{\sigma_X^2}\right) \propto \exp-\frac{1}{2}\sum_{i=1}^{n}\left(\frac{(x_i-\mu)^2}{\sigma_X^2}\right)$$

$$\propto \exp-\frac{1}{2\sigma_X^2}\left(n\mu^2-2\mu\sum_{i=1}^{n}x_i\right) \propto \exp-\frac{n}{2\sigma_X^2}\left(\mu-\frac{1}{n}\sum_{i=1}^{n}x_i\right)^2$$

Is a normal distribution $\phi(\mu;\nu,\tau^2)$ with parameters: $\nu=\frac{1}{n}\sum_{i=1}^{n}x_i\,,\tau^2=\frac{\sigma_X^2}{n}$

$$p(x_i|\sigma_R^2,\sigma_X^2,\mu,\boldsymbol{x}_{-i},\boldsymbol{y}) \propto \exp-\frac{1}{2}\frac{(x_i-\mu)^2}{\sigma_X^2}\prod_{j=1}^{2}\exp-\frac{1}{2}\left(\frac{\left(y_{ij}-x_i\right)^2}{\sigma_R^2}\right)$$

$$\propto \exp-\frac{1}{2}\left(\left(\frac{1}{\sigma_X^2}+\frac{2}{\sigma_R^2}\right)x_i^2-2x_i\left(\frac{\mu}{\sigma_X^2}+\frac{y_{i1}+y_{i2}}{\sigma_R^2}\right)\right)$$

Is a normal distribution $\phi(x_i;\nu,\tau^2)$ with parameters: $\nu=\left(\frac{\mu}{\sigma_X^2}+\frac{y_{i1}+y_{i2}}{\sigma_R^2}\right)\tau^2\,,\tau^2=\left(\frac{1}{\sigma_X^2}+\frac{2}{\sigma_R^2}\right)^{-1}$