UNIVERSITY OF OSLO

Faculty of mathematics and natural sciences

Exam in:	STK4051 — Computational statistics Suggested solutions
Day of examination:	Tuesday May 28 2019.
Examination hours:	14.30 – 18.30.
This problem set consists of 5 pages.	
Appendices:	None
Permitted aids:	None

Please make sure that your copy of the problem set is complete before you attempt to answer anything.

Problem 1

(a) For observations that are We have that are truncated at zero, we have

$$\Pr(y_i = 0) = \Pr(x_i \leq 0) = \Phi(0; \mu, \sigma)$$

giving

$$l(\boldsymbol{\theta}) = -\frac{n_1}{2}\log(2\pi) - \frac{n_1}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i - \mu)^2 + n_2 \log(\Phi(0; \mu, \sigma))$$

Since we have the constraint $\sigma^2 > 0$, reparametrizing to $\rho = \log(\sigma^2)$ gives a non-constrained optimization problem.

(b) Newton-like methods:

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - [\boldsymbol{M}^t]^{-1} l'(\boldsymbol{\theta}^t)$$

where $\boldsymbol{M}^t$ is a matrix approximating the Hessian. Ascent algorithms use simple forms of $\boldsymbol{M}^t$, e.g. the identity matrix. This method usually converge fast (if it converge) but require computation of gradients.

Nelder-Mead: With $\boldsymbol{\theta}$ 2-dimensional, we start with 3 values of $\boldsymbol{\theta}$. These 3 values are dynamically altered by replacing the worst value with a better one, defined through a search line going through the worst value and the average of the other values. The worst value is then updated to a better (best?) value along this line. The algorithm is performing these steps iteratively until some stopping criterion is achieved. This method does not need the derivatives and is typically quite robust.

(Continued on page 2.)

(c) We have the complete -loglikelihood

$$l_c(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$$

We then get

$$\begin{aligned}
Q(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}E[(x_i-\mu)^2|y_i,\boldsymbol{\theta}^t] \\
=& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i-\mu)^2 - \\
& \frac{n_2}{2\sigma^2}E[(x_i-\mu)^2|x_i<0,\boldsymbol{\theta}^t] \\
=& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i-\mu)^2 - \\
& \frac{n_2}{2\sigma^2}[E[x_i^2|x_i<0,\boldsymbol{\theta}^t] - 2E[x_i|x_i<0,\boldsymbol{\theta}^t]\mu+\mu^2] \\
=& -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n_1}(y_i-\mu)^2 - \\
& \frac{n_2}{2\sigma^2}[m_2(\mu^t,(\sigma^2)^t) - 2m_1(\mu^t,(\sigma^2)^t)\mu+\mu^2]
\end{aligned}$$

Taking derivatives, we get

$$\frac{\partial}{\partial\mu}Q(\boldsymbol{\theta},\boldsymbol{\theta}^t) = \frac{1}{\sigma^2}\sum_{i=1}^{n_1}(y_i-\mu) + \frac{n_2}{\sigma^2}[m_1(\mu^t,(\sigma^2)^t)-\mu]$$

giving

$$\mu^{t+1} = \frac{\sum_{i=1}^{n_1}y_i + n_2 m_1(\mu^t,(\sigma^2)^t)}{n}$$

Further,

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}Q(\boldsymbol{\theta},\boldsymbol{\theta}^t) =& -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n_1}(y_i-\mu)^2 + \\
& \frac{n_2}{2\sigma^4}[m_2(\mu^t,(\sigma^2)^t) - 2m_1(\mu^t,(\sigma^2)^t)\mu+\mu^2]
\end{aligned}$$

giving

$$(\sigma^2)^{t+1} = \frac{1}{n}[\sum_{i=1}^{n_1}(y_i-\mu^{t+1})^2 + n_2[m_2(\mu^t,(\sigma^2)^t) - 2m_1(\mu^t,(\sigma^2)^t)\mu^{t+1} + (\mu^{t+1})^2]]$$

(Continued on page 3.)

(d) We can obtain Monte Carlo estimates of the m_p's by simulating $x_j \sim N(\mu^t, (\sigma^2)^t), j = 1, ..., N$, discard those values for which $x_j > 0$ and then take the means of x_j and x_j^2 of the remaining samples.

Due to the stochasticity of the Monte Carlo estimates, the properties of the EM algorithm is now violated giving jumps up and down without stabilization.

Problem 2

(a) We have

$$\begin{aligned} P(z|x) &= \int_y P_1(y|x)P_2(z|y)dy \\ \int_x \pi(x)P(z|x)dx &= \int_x \pi(x) \int_y P_1(y|x)P_2(z|y)dydx \\ &= \int_y P_2(z|y) \int_x \pi(x)P_1(y|x)dxdy \\ &= \int_y P_2(z|y)\pi(y)dy \\ &= \pi(z) \end{aligned}$$

showing the result.

(b) Assume the current value is $\boldsymbol{x} = (x_1, x_2)$ while we simulate a new value $\boldsymbol{x}^* = (x_1^*, x_2)$ by P_1. Then

$$\begin{aligned} \pi(\boldsymbol{x})P_1(\boldsymbol{x}^*|\boldsymbol{x}) &= \pi(x_1, x_2)\pi(x_1^*|x_2) = \pi(x_2)\pi(x_1|x_2)\pi(x_1^*|x_2) \\ &= \pi(x_1|x_2)\pi(x_1^*, x_2) = \pi(\boldsymbol{x}^*)P_1(\boldsymbol{x}|\boldsymbol{x}^*) \end{aligned}$$

showing that detailed balance is obained for P_1. We get similar results for P_2.

The systematic Gibbs sampler then satisfies the general properties of P_1 and P_2 considered in (a) and this shows that the systematic scan Gibbs sampler keeps the target distribution invariant.

(c) We have in this case that

$$\begin{aligned} R &= \frac{\pi(\boldsymbol{x}^*)\pi(x_1|x_2)}{\pi(\boldsymbol{x})\pi(x_1^*|x_2)} = \frac{\pi(x_1^*, x_2)\pi(x_1|x_2)}{\pi(x_1, x_2)\pi(x_1^*|x_2)} \\ &= \frac{\pi(x_1^*|x_2)\pi(x_2)\pi(x_1|x_2)}{\pi(x_1|x_2)\pi(x_2)\pi(x_1^*|x_2)} = 1 \end{aligned}$$

showing that with this proposal we always accept and actually end up with a Gibbs sampler step.

Gibbs sampling is therefore a special case of Metropolis-Hastings.

(Continued on page 4.)

(d) We have that

$$\begin{aligned}
\pi(\mu|\boldsymbol{x},\tau) \propto& \pi(\mu,\tau)p(\boldsymbol{x}|\mu,\tau)\\
\propto& \prod_{i=1}^{n} e^{-0.5\tau(x_i-\mu)^2}\\
\propto& e^{-0.5\tau(\sum_{i=1}^{n} x_i^2 - 2\mu\sum_{i=1}^{n} x_i + n\mu^2)}\\
\propto& e^{-0.5n\tau(\mu^2-\mu\bar{x})}\\
\propto& e^{-0.5n\tau(\mu-\bar{x})^2} \propto N(\bar{x},(n\tau)^{-1})
\end{aligned}$$

Similarly

$$\begin{aligned}
\pi(\tau|\boldsymbol{x},\mu) \propto& \pi(\mu,\tau)p(\boldsymbol{x}|\mu,\tau)\\
\propto& \tau^{\alpha-1}e^{-\beta\tau}\prod_{i=1}^{n}\tau^{1/2}e^{-0.5\tau(x_i-\mu)^2}\\
\propto& \tau^{\alpha+0.5n-1}e^{-\tau[\beta+\sum_{i=1}^{n}(x_i-\mu)^2]}\\
\propto& \text{Gamma}(\alpha+0.5n,\beta+\sum_{i=1}^{n}(x_i-\mu)^2)
\end{aligned}$$

showing that we can switch between sampling μ from a Gaussian distribution and τ from a Gamma distribution.

Problem 3

(a) In the stochastic gradient algorithm we replace $\nabla g(\boldsymbol{\beta})$ by an unbiased estimate. By doing so, we also need to change the learning rate to α_t which needs spesific properties in order for the algorithm to converge.

Due to that an estimate of the gradient can be much cheaper to calculate, we can benefit from this in this setting due to the large number of observations.

(b) The conditions needed are

$$\alpha_t > 0 \tag{A-1}$$

$$\sum_{t=2}^{\infty}\frac{\alpha_t}{\alpha_1+\cdots+\alpha_{t-1}} = \infty \tag{A-2}$$

$$\sum_{t=1}^{\infty}\alpha_t^2 < \infty \tag{A-3}$$

The first is reasonable since we want to move in the right direction. The second is reasonable since we do not want to move in too small steps, while the last is reasonable since if the learning rates becomes too large, the variance of the cumulative gradients will explode.

(Continued on page 5.)

(c) By a minibatch we mean exactly to draw a random subsample from $\{1, ..., n\}$ of size m. Larger m decrease variability but increase computational effort. Since the variability is largest in the first plot, this should correspond to the smallest m, with increasing order.