SAMPLE EXAM 3 SOLUTIONS

• This is a CLOSED BOOK exam. You may use one page, both sides, of handwritten notes.

• There are a total of 100 points in the exam. Plan your work accordingly.

• You must SHOW YOUR WORK to get full credit.

<table>
<thead>
<tr>
<th>Problem</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

Name: ____________________________________________
Problem 1  (20 points)

An RNN has output $\hat{y}(t)$ defined for $0 \leq t < \infty$, input $x(t)$, error $E_t$, network weights $a_t$ and $b_t$, and scalar nonlinearity $g(\cdot)$ related by

\[ \hat{y}(t) = g(a_t \hat{y}(t-1) + b_t x(t)) \]  
\[ E_t = \frac{1}{2} (y(t) - \hat{y}(t))^2 \]  
\[ a_0 = b_0 = 0 \]

\[ a_{t+1} = a_t - 0.02 \sum_{k=0}^{t} \frac{\partial E_t}{\partial a_k} \]  

(a) Prove that, even if $E_t$ is bounded, $\sum_{k=0}^{t} \frac{\partial E_t}{\partial a_k}$ might grow without bound.

**Solution:**

For convenience, define $g'(t) = \frac{\partial g(a_t \hat{y}(t-1) + b_t x(t))}{\partial a_t \hat{y}(t-1) + b_t x(t)}$. Then

\[ \sum_{k=0}^{t} \frac{\partial E_t}{\partial a_k} = (y(t) - \hat{y}(t)) \sum_{k=0}^{t} \frac{\partial \hat{y}(t)}{\partial a_k} \]

\[ = (y(t) - \hat{y}(t)) \sum_{k=0}^{t} \hat{y}(t-1-k) g'(t) \prod_{\ell=1}^{k} a_{t-\ell} g'(t-\ell) \]

\[ \to \infty \text{ as } t \to \infty \]

where the last line holds if the geometric mean of $|a_{t-\ell} g'(t-\ell)|$ is greater than 1.

(b) Modify the update equation to

\[ a_{t+1} = \max \left( -a_{MAX}, \min \left( a_{MAX}, a_t - 0.02 \sum_{k=0}^{t} \frac{\partial E_t}{\partial a_k} \right) \right) \]

Find sufficient conditions on $a_{MAX}$ and $g(\cdot)$ such that $\sum_{k=0}^{t} \frac{\partial E_t}{\partial a_k}$ is guaranteed to remain bounded as $t \to \infty$.

**Solution:** It is sufficient if we can guarantee that $|a_{t-\ell} g'(t-\ell)| < 1$ at all times. This is achieved if

\[ a_{MAX} < \frac{1}{\max_x |\frac{dg}{dx}|} \]

Problem 2  (10 points)

Consider a two-layer RNN, one node per layer, with input $x(t)$, hidden layer $h(t)$, output $\hat{y}(t)$, error $E_t$, scalar nonlinearity $g(\cdot)$, and coefficients $a, b, \alpha, \beta$ related by

\[ h(t) = g(ah(t-1) + bx(t)) \]
\[ \hat{y}(t) = (\alpha \hat{y}(t-1) + \beta h(t)) \]
\[ E_t = y(t) \log(\hat{y}(t)) + (1 - y(t)) \log(1 - \hat{y}(t)) \]

Find $\frac{\partial E_t}{\partial a}$ and $\frac{\partial E_t}{\partial b}$. 

---

NAME:__________________________  Sample Exam 3  Page 2
Solution:

For convenience, define \( g'(t) = \frac{\partial g(ah(t-1) + bx(t))}{\partial ah(t-1) + bx(t)} \). Then

\[
\frac{\partial E_t}{\partial \beta} = \left( \frac{y(t)}{\hat{y}(t)} - \frac{1 - y(t)}{1 - \hat{y}(t)} \right) \beta \frac{\partial h(t)}{\partial b} + \alpha \frac{\partial \hat{y}(t-1)}{\partial \beta} \]

\[
= \left( \frac{y(t)}{\hat{y}(t)} - \frac{1 - y(t)}{1 - \hat{y}(t)} \right) \beta \sum_{k=0}^{t} \alpha^k h(t-k)
\]

\[
= \left( \frac{y(t)}{\hat{y}(t)} - \frac{1 - y(t)}{1 - \hat{y}(t)} \right) \beta \sum_{k=0}^{t} \alpha^k \frac{g'(t-k)}{g'(t-k-m)} x(t-k) + a \frac{\partial h(t-k-1)}{\partial b}
\]

Problem 3  (20 points)

A simplified LSTM with inputs \( x(t) \) is defined by

\[
c(t) = i(t)x(t) + m(t)c(t-1) \quad \text{(8)}
\]

\[
i(t) = u(w^i x(t) + b^i) \quad \text{(9)}
\]

\[
m(t) = u(w^m x(t) + b^m) \quad \text{(10)}
\]

where \( u(\cdot) \) is the unit step function, defined as \( u(z) = \frac{1}{2} (\text{sign}(z) + 1) \).

(a) Choose \( w^i, b^i, w^m \) and \( b^m \) so that

\[
c(t) = \begin{cases} 
  x(t) & x(t) > 2 \\
  c(t-1) & x(t) < 2
\end{cases}
\]

\textbf{Solution:} This is solved by \( w^i = 1, w^m = -1, b^i = -2, b^m = 2 \).

(b) Suppose \( x(t) = s(t) + v(t) \), where \( s(t) \) is the desired signal,

\[
s(t) = \begin{cases} 
  10 & t = 0 \\
  0 & \text{otherwise}
\end{cases}
\]

and \( v(t) \) is a random noise process with distribution given by

\[
\Phi(z) = \Pr \{ v(t) \leq z \} = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}u^2} du
\]
Define the “memory” of the network you designed in part (a) to be the expected number of time steps for which \( c(t) = x(0) \). Find the memory of your network, as a function of \( \Phi(z) \).

**Solution:** Let \( T \) be the number of time steps for which \( c(t) = x(0) \); the “memory” is \( E[T] \).

- \( T = 0 \) if \( 10 + v(0) < 2 \), i.e., if \( v(0) < -8 \); this happens with probability \( p_0 = \Phi(-8) \).
- \( T = 1 \) if \( v(0) > -8 \) and \( v(1) > 2 \); this happens with probability \( p_0 p_1 \) where \( p_1 = 1 - \Phi(2) \).
- \( T = t \) with probability \( p_0(1 - p_1)^{t-1} p_1 \).

The expected value \( E[T] \) is therefore

\[
E[T] = \sum_{t=1}^{\infty} t(1 - p_0)(1 - p_1)^{t-1} p_1
\]

We can solve this sum by noting that

\[
\sum_{t=0}^{\infty} (1 - p_1)^t = \frac{1}{p_1}
\]

Differentiating both sides by \( p_1 \), we find that

\[
\sum_{t=0}^{\infty} t(1 - p_1)^{t-1} = \frac{1}{p_1^2}
\]

Therefore

\[
E[T] = \frac{1 - p_0}{p_1} = \frac{1 - \Phi(-8)}{1 - \Phi(2)}
\]

**Problem 4 (25 points)**

Define \( c_t = i \) to be the event that the \( i^{th} \) coin is flipped at time \( t \), where \( 1 \leq i \leq 3 \). The possible observations are \( x_t \in \{H, T\} \). Two of the coins are unfair; the heads probabilities of the three coins are given by \( p(x_t = H|c_t = i) = i/4 \). Coin \( c_1 = 1 \) is always the first one flipped. After each coin flip, the coin is changed with probability \( 1/2 \); if the coin is changed, both of the other coins are equally likely.

(a) What is \( p(c_1 = 1, c_2 = 1, c_3 = 2) \)?

**Solution:**

\[
p(c_1 = 1, c_2 = 1, c_3 = 2) = \pi_1 a_{11} a_{12} = (1)(\frac{1}{2})(\frac{1}{4}) = \frac{1}{8}
\]

(b) What is the probability of getting three heads in a row?

**Solution:** Define \( \alpha_1(i) = \pi_i b_1(x_1) \). Then

\[
\alpha_1(k) = \begin{cases} \frac{1}{4} & k = 1 \\ 0 & \text{otherwise} \end{cases}
\]
Define \( \alpha_t(j) = \sum_i \alpha_{t-1}(i) a_{ij} b_j(x_t) \). Then
\[
\alpha_2(i) = \begin{cases} 
\left( \frac{1}{4} \right) \left( \frac{1}{2} \right) \left( \frac{1}{4} \right) = \frac{1}{32} & i = 1 \\
\left( \frac{1}{4} \right) \left( \frac{1}{2} \right) \left( \frac{3}{4} \right) = \frac{3}{32} & i = 2 \\
\left( \frac{1}{4} \right) \left( \frac{1}{2} \right) \left( \frac{3}{4} \right) = \frac{3}{32} & i = 3
\end{cases}
\]
and
\[
\alpha_3(j) = \begin{cases} 
\left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) = \frac{9}{1536} & j = 1 \\
\left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) = \frac{9}{1536} & j = 2 \\
\left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) + \left( \frac{1}{64} \right) \left( \frac{1}{32} \right) \left( \frac{1}{2} \right) \left( \frac{1}{2} \right) = \frac{9}{1536} & j = 3
\end{cases}
\]
Adding them all up, we get \( p(x_1 = H, x_2 = H, x_3 = H) = \frac{57}{1024} \).

(c) Suppose you observe two heads in a row. You know that the first flip was coin \( c_1 = 1 \). Which coin was most probably the second one flipped?

**Solution:**
\[
p(c_2 = 1, HH) = \left( \frac{1}{4} \right) \left( \frac{1}{2} \right) \left( \frac{1}{4} \right) = \frac{1}{32} \\
p(c_2 = 2, HH) = \left( \frac{1}{4} \right) \left( \frac{1}{4} \right) \left( \frac{1}{2} \right) = \frac{1}{32} \\
p(c_2 = 3, HH) = \left( \frac{1}{4} \right) \left( \frac{1}{4} \right) \left( \frac{3}{4} \right) = \frac{3}{64}
\]
So the most likely second coin is \( c_2 = 3 \).

**Problem 5** (15 points)

Suppose that an HMM has 6 states with transition probabilities arranged into a matrix \( A \), whose \((i, j)^{th}\) element is \( a_{ij} = \Pr \{ s_t = j | s_{t-1} = i \} \). The \( a_{ij} \) are initialized randomly, but because of a programming error, \( a_{34} \) is accidentally set to \( a_{34} = 0 \). New transition probabilities \( \hat{a}_{ij} \) are estimated as
\[
\hat{a}_{ij} = \frac{\sum_{t=1}^{T} x_t(i,j)}{\sum_{t=1}^{T} \sum_{j=1}^{6} x_t(i,j)} 
\]
where \( X = [x_1, \ldots, x_T] \) is a sequence of observations whose details you do not need to know.

Prove that, under these circumstances, \( \hat{a}_{34} = 0 \).

**Solution:**
\[
\xi_t(3, 4) = \frac{p(X, s_{t-1} = 3, s_t = 4 | A)}{p(X | A)} = \alpha_{t-1}(3) a_{34} b_4(x_t) \beta_t(4) = 0
\]
Therefore \( \sum_{t=1}^{T} \xi_t(3, 4) = 0 \), therefore \( \hat{a}_{34} = 0 \).

**Problem 6** (10 points)

A NN-HMM hybrid has known observations \( X = [x_1, \ldots, x_T] \), and unknown state sequence \( S = [s_1, \ldots, s_T] \). Its initialization probabilities are \( \pi_t = \Pr \{ s_1 = i \} \), and its transition probabilities are \( a_{ij} = \Pr \{ s_t = j | s_{t-1} = i \} \). Its observation probabilities are computed by the softmax
layer of a neural network; they are $b_j(x_t) = \Pr \{ s_t = j \mid x_t \} / \Pr \{ s_t = j \}$. Consider the following algorithm:

$$
\alpha_1(i) = \pi_i b_i(x_1), \quad 1 \leq i \leq N \tag{14}
$$

$$
\alpha_{t+1}(j) = \sum_{i=1}^{N} \alpha_t(i) a_{ij} b_j(x_t), \quad 1 \leq i, j \leq N, \quad 1 \leq t \leq T - 1 \tag{15}
$$

$$
P_{final} = \sum_{j=1}^{N} \alpha_T(j) \tag{16}
$$

Express $P_{final} = \ldots$, where the right-hand side of the equation contains some sort of arithmetic combination of joint, conditional, and marginal probability mass functions of the random variables $x_t$. No other variables should appear on the right-hand side.

**Solution:** First let’s use the definition of conditional probability to rewrite $b_j(x_t)$ in a more useful form:

$$
b_j(x_t) = \frac{\Pr \{ s_t = j \mid x_t \}}{\Pr \{ s_t = j \}} = \frac{\Pr \{ x_t \mid s_t = j \} p(x_t)}{p(x_t)}
$$

Then

$$
\alpha_1(i) = \pi_i b_i(x_1) = \frac{p(x_1, s_1 = i)}{p(x_1)}
$$

Continuing to $t = 2$, we have

$$
\alpha_2(j) = \sum_{i=1}^{N} \frac{p(x_1, s_1 = i)}{p(x_1)} a_{ij} \frac{p(x_2 | s_2 = j)}{p(x_2)}
$$

$$
= \sum_{i=1}^{N} \frac{p(x_1, s_1 = i, s_2 = j, x_2)}{p(x_1)p(x_2)}
$$

$$
= \frac{p(x_1, x_2, s_2 = j)}{p(x_1)p(x_2)}
$$

Continuing, we find that

$$
P_{final} = \frac{p(x_1, x_2, \ldots, x_T)}{p(x_1)p(x_2) \ldots p(x_T)}$$