1. What is the difference between backpropagation algorithm and backpropagation through time (BPTT) algorithm ?

   A. There is no difference.

   B. Unlike backpropagation, in BPTT we add the gradients for corresponding weight for each time step.

   C. Unlike backpropagation, in BPTT we subtract the gradients for corresponding weight for each time step.

Solution: **Option B** is the correct answer.

2. What approach is taken to deal with the problem of Exploding Gradients in Recurrent Neural Networks?

    A. Gradient clipping

    B. Using modified architectures like LSTMs and GRUs

    C. Using dropout

Solution: **Option A** is the correct option.

3. In the context of the state equations of LSTM, we have seen that $h_t = o_t \odot \sigma(s_t)$ where $h_t, o_t, s_t \in \mathbb{R}^n$. What is the derivative of h_t w.r.t. s_t?

   A. Vector

   B. Tensor

   C. Matrix

Solution: **Option C** is the correct answer.
The derivative will yield a diagonal square matrix.

4. Continuing the previous question, how many non-zero entries does the derivative of h_t w.r.t. s_t have?

   A. No non-zero entries

   B. n

   C. $n^2 - n$

Solution: **Option B** is the correct answer.
The derivative will yield a diagonal square matrix. The only non-zero elements will be the ones on the diagonal.

5. In the context of LSTMs, the gradient of $\mathcal{L}_t(\theta)$ w.r.t θ_i vanishes when

   A. the gradients flowing through at least one path from $\mathcal{L}_t(\theta)$ to θ_i vanishes.

   B. the gradients flowing through each and every path from $\mathcal{L}_t(\theta)$ to θ_i vanishes.

Solution: **Option B** is the correct answer

6. Which of the following options represent the full set of equations for GRU gates where s_t represents the state of the GRU and h_t refers to the intermediate output?

   A. $\mathrm{o}_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$

   $i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$

   B. $\mathrm{o}_t = \sigma(W_o s_{t-1} + U_o x_t + b_o)$

   $i_t = \sigma(W_i s_{t-1} + U_i x_t + b_i)$

Solution: Option B is the correct answer.

7. Consider a GRU where the input $x \in \mathbb{R}^m$ and the state of GRU $s \in \mathbb{R}^n$ at any time step t. What is the total number of parameters in this GRU ?

    A. $n^2 + nm + 2n$

    B. $3 \times (n^2 + nm + n)$

    C. $n + 3 \times (n^2 + nm + n)$

    D. $4 \times (n^2 + nm + n)$

Solution: From the equations of GRU, we know that it has the parameters namely, $W_o, W_i, W, U_o, U_i, U, b_o, b_i, b$ where the dimensions of $W's \in \mathbb{R}^{n \times n}$, $U's \in \mathbb{R}^{m \times n}$ and b's $\in \mathbb{R}^n$. Therefore, **Option B** is the correct answer.

8. Consider the following statements in the context of LSTMs:

    1. During forward propagation, the gates control the flow of information.
    2. During backward propagation, the gates control the flow of gradients.

    Which of the following option is correct ?

    A. Statement 1 is True and Statement 2 is False.

    B. Statement 2 is True and Statement 1 is False.

    C. Both are False.

    D. Both are True.

Solution: **Option D** is the correct answer.

9. Consider the RNN with the following equations:

$$s_t = \sigma(Ux + Ws_{t-1} + b)$$
$$y_t = \mathcal{O}(Vs_t + c)$$

where s_t is the state of the network at timestep t and the parameters W,U,V,b,c are shared across timesteps. The loss $\mathscr{L}_t(\theta)$ is defined as :

$$\mathscr{L}_t(\theta) = -\log(y_{tc})$$

where y_{tc} is the predicted probability of true output at time-step t. Given the above RNN, find $\frac{\partial \mathscr{L}_t(\theta)}{\partial s_t}$ at $t = 4$.

A. $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -\frac{\mathcal{O}(Vs_4+c)}{\mathcal{O}'(Vs_4+c)}$

B. $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\frac{\mathcal{O}(Vs_4+c)}{\mathcal{O}'(Vs_4+c)}$

C. $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\frac{\mathcal{O}'(Vs_4+c)}{\mathcal{O}(Vs_4+c)}$

D. $\frac{\partial \mathscr{L}_4(\theta)}{\partial s_4} = -V\mathcal{O}'(Vs_4 + c)$

Solution: Option C is the correct answer.

10. Considering the same RNN setup as defined in the previous question, find $\frac{\partial \mathscr{L}(\theta)}{\partial V}$.

A. $\frac{\partial \mathscr{L}(\theta)}{\partial V} = -s_t \frac{\mathcal{O}(V s_t + c)}{\mathcal{O}'(V s_t + c)}$

B. $\frac{\partial \mathscr{L}(\theta)}{\partial V} = -s_t \frac{\mathcal{O}'(V s_t + c)}{\mathcal{O}(V s_t + c)}$

C. $\frac{\partial \mathscr{L}(\theta)}{\partial V} = \sum_{t=1}^{T} -s_t \frac{\mathcal{O}(V s_t + c)}{\mathcal{O}'(V s_t + c)}$

D. $\frac{\partial \mathscr{L}(\theta)}{\partial V} = \sum_{t=1}^{T} -s_t \frac{\mathcal{O}'(V s_t + c)}{\mathcal{O}(V s_t + c)}$

Solution: Option D is the correct answer.