1. Which of the following are true about hyperparameter search? 1 / 1 point

- ○ Choosing values in a grid for the hyperparameters is better when the number of hyperparameters to tune is high since it provides a more ordered way to search.
- ◉ Choosing random values for the hyperparameters is convenient since we might not know in advance which hyperparameters are more important for the problem at hand.
- ○ When sampling from a grid, the number of values for each hyperparameter is larger than when using random values.
- ○ When using random values for the hyperparameters they must be always uniformly distributed.

✓ **Correct**
Correct. Different problems might be more sensitive to different hyperparameters.

2. If it is only possible to tune two parameters from the following due to limited computational resources. Which two would you choose? 1 / 1 point

- ☐ β_1, β_2 in Adam.
- ☑ The β parameter of the momentum in gradient descent.

✓ **Correct**
Correct. This hyperparameter can increase the speed of convergence of the training, thus is worth tuning.

- ☐ ϵ in Adam.
- ☑ α

✓ **Correct**
Correct. This might be the hyperparameter that most impacts the results of a model.

3. During hyperparameter search, whether you try to babysit one model ("Panda" strategy) or train a lot of models in parallel ("Caviar") is largely determined by: 1 / 1 point

- ◉ The amount of computational power you can access
- ○ The number of hyperparameters you have to tune
- ○ Whether you use batch or mini-batch optimization
- ○ The presence of local minima (and saddle points) in your neural network

✓ **Correct**

4. If you think β (hyperparameter for momentum) is between 0.9 and 0.99, which of the following is the recommended way to sample a value for beta? 1 / 1 point

- r = np.random.rand() beta = 1-10**(- r + 1)
- r = np.random.rand() beta = r*0.9 + 0.09
- r = np.random.rand() beta = 1-10**(- r - 1)
- r = np.random.rand() beta = r*0.09 + 0.9

Correct

5. Once good values of hyperparameters have been found, those values should be changed if new data is added or a change in computational power occurs. True/False? 1 / 1 point

- False
- True

Correct
Correct. The choice of some hyperparameters such as the batch size depends on conditions such as hardware and quantity of data.

6. When using batch normalization it is OK to drop the parameter $W^{[l]}$ from the forward propagation since it will be subtracted out when we compute $\tilde{z}^{[l]} = \gamma z^{[l]}_{\text{normalize}} + \beta^{[l]}$. True/False? 1 / 1 point

- True
- False

Correct
Correct. The parameter $W^{[l]}$ doesn't get subtracted during the batch normalization process, although it gets re-scaled.

7. Which of the following are true about batch normalization? 1 / 1 point

- ○ There is a global value of γ and β that is used for all the hidden layers where batch normalization is used.
- ◉ One intuition behind why batch normalization works is that it helps reduce the internal covariance.
- ○ The parameters β and γ of batch normalization can't be trained using Adam or RMS prop.
- ○ The parameter ϵ in the batch normalization formula is used to accelerate the convergence of the model.

Correct
Yes. Internal covariance is a name to express that there has been a change in the distribution of the activations. Since after each iteration of gradient descent the parameters of a layer change, we might think that the activations suffer from covariance shift.

1 / 1 point

8. Which of the following are true about batch normalization?

- ☐ $\beta^{[l]}$ and $\gamma^{[l]}$ are hyperparameters that must be tuned by random sampling in a logarithmic scale.
- ☑ When using batch normalization we introduce two new parameters $\gamma^{[l]}, \beta^{[l]}$ that must be "learned" or trained.

Correct
Correct. Batch normalization uses two parameters β and γ to compute $\tilde{z}^{(i)} = \beta z_{norm}^{(i)} + \gamma$.

- ☐ $z_{norm}^{(i)} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2}}$.
- ☑ The parameters $\gamma^{[l]}$ and $\beta^{[l]}$ set the variance and mean of $\tilde{z}^{[l]}$.

Correct
Correct. When applying the linear transformation $\tilde{z}^{(l)} = \beta^{[l]} z_{norm}^{(l)} + \gamma^{[l]}$ we set the variance and mean of $\tilde{z}^{[l]}$.

9. A neural network is trained with Batch Norm. At test time, to evaluate the neural network on a new example you should perform the normalization using μ and σ^2 estimated using an exponentially weighted average across mini-batches seen during training. True/false? 1 / 1 point

- ◉ True
- ○ False

Correct
Correct. This is a good practice to estimate the μ and σ^2 to use since at test time we might not be

- [ ] $\beta^{[l]}$ and $\gamma^{[l]}$ are hyperparameters that must be tuned by random sampling in a logarithmic scale.
- [x] When using batch normalization we introduce two new parameters $\gamma^{[l]}$, $\beta^{[l]}$ that must be "learned" or trained.

Correct
Correct. Batch normalization uses two parameters β and γ to compute $\tilde{z}^{(i)} = \beta z^{(i)}_{norm} + \gamma$.

- [ ] $z^{(i)}_{norm} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2}}$.
- [x] The parameters $\gamma^{[l]}$ and $\beta^{[l]}$ set the variance and mean of $\tilde{z}^{[l]}$.

Correct
Correct. When applying the linear transformation $\tilde{z}^{(l)} = \beta^{[l]} z^{(l)}_{norm} + \gamma^{[l]}$ we set the variance and mean of $\tilde{z}^{[l]}$.

9. A neural network is trained with Batch Norm. At test time, to evaluate the neural network on a new example you should perform the normalization using μ and σ^2 estimated using an exponentially weighted average across mini-batches seen during training. True/false? 1 / 1 point

- (•) True
- ( ) False

Correct
Correct. This is a good practice to estimate the μ and σ^2 to use since at test time we might not be predicting over a batch of the same size, or it might even be a single example, thus using the μ and σ^2 of a single sample doesn't make sense.

10. Which of the following are some recommended criteria to choose a deep learning framework? 1 / 1 point

- ( ) It must run exclusively on cloud services, to ensure its robustness.
- ( ) It must be implemented in C to be faster.
- ( ) It must use Python as the primary language.
- (•) Running speed.

Correct
Correct. The running speed is a major factor, especially when working with large datasets.