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

- When using random values for the hyperparameters they must be always uniformly distributed.
- (selected) 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.
- 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.
- When sampling from a grid, the number of values for each hyperparameter is larger than when using random values.

Expand

✓ **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**

- (selected) 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.
- (selected) α
  - ✓ **Correct**
    Correct. This might be the hyperparameter that most impacts the results of a model.
- ☐ β_1, β_2 in Adam.

Expand

✓ **Correct**
Great, you got all the right answers.

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**

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

Expand

✓ **Correct**

4. Knowing that the hyperparameter α should be in the range of 0.001 and 1.0. Which of the following is the recommended way to sample a value for α? **0 / 1 point**

- r = 4*np.random.rand()
  alpha = 10**r
- (selected) r = -5*np.random.rand()
  alpha = 10**r
- r = -3*np.random.rand()
  alpha = 10**r
- r = np.random.rand()
  alpha = 0.001 + r*0.999

Expand

✗ **Incorrect**
No. This will generate a random value between 10^{-5} and 10^{0}.

5. Finding new values for the hyperparameters, once we have found good ones for a model, should only be done if new hardware or computational power is acquired. True/False? **1 / 1 point**

- True
- (selected) False

Expand

✓ **Correct**
Correct. As the data changes for the model, it might be beneficial to tune some of the hyperparameters again.

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
- (selected) False

Expand

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

7. In the normalization formula $z^{(i)}_{\text{norm}} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2 + \epsilon}}$, why do we use epsilon? **1 / 1 point**

- In case μ is too small
- To have a more accurate normalization
- To speed up convergence
- (selected) To avoid division by zero

Expand

✓ **Correct**

8. Which of the following is true about batch normalization? **1 / 1 point**

- (selected) The parameters $\gamma^{[l]}$ and $\beta^{[l]}$ set the variance and mean of $\tilde{z}^{[l]}$.
- The optimal values to use for γ and β are $\gamma = \sqrt{\sigma^2 + \epsilon}$ and $\beta = \mu$.
- $z^{(i)}_{\text{norm}} = \frac{z^{(i)} - \mu}{\sqrt{\sigma^2}}$.
- The parameters $\gamma^{[l]}$ and $\beta^{[l]}$ can be learned only using plain gradient descent.

Expand

✓ **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. After training a neural network with Batch Norm, at test time, to evaluate the neural network on a new example you should: **1 / 1 point**

- If you implemented Batch Norm on mini-batches of (say) 256 examples, then to evaluate on one test example, duplicate that example 256 times so that you're working with a mini-batch the same size as during training.
- Use the most recent mini-batch's value of μ and σ^2 to perform the needed normalizations.
- Skip the step where you normalize using μ and σ^2 since a single test example cannot be normalized.
- (selected) Perform the needed normalizations, use μ and σ^2 estimated using an exponentially weighted average across mini-batches seen during training.

Expand

✓ **Correct**

10. Which of these statements about deep learning programming frameworks are true? (Check all that apply) **1 / 1 point**

- (selected) A programming framework allows you to code up deep learning algorithms with typically fewer lines of code than a lower-level language such as Python.
  - ✓ **Correct**
- ☐ Deep learning programming frameworks require cloud-based machines to run.
- (selected) Even if a project is currently open source, good governance of the project helps ensure that it remains open even in the long term, rather than become closed or modified to benefit only one company.
  - ✓ **Correct**

Expand

✓ **Correct**
Great, you got all the right answers.