# Different Types of Activation Units – Quiz

1. Select all (one or more) correct statements about a neural network (MLP) with the *identity* activation function $\sigma(x) = x$ trained with a mean squared error (MSE) loss function:
(a) It expresses the same family of functions as linear regression does.
(b) It expresses the same family of functions as ReLU nets with the same depth and width.
(c) Given a training dataset, if the loss surface of linear regression has a unique global minimum, then the loss surface of this neural network will also have a unique global minimum.

The correct answer is (a).
Neural networks with linear activations are equivalent to linear regression models, only in the sense of representation rather than optimization. The reason why (a) is true and (b) is false has already been explained in the assignment. (c) is false, and a simple example is $f(x) = abx$, fitted with $(1, 1)$; the loss will be $(ab - 1)^2$, where there are infinite number of global minimums satisfying $ab = 1$; however, for its 1-d linear regression, the loss is $(w - 1)^2$, where there is one unique global minimum $w = 1$.

2. Explain what **gradient vanishing** is and how it could come about in neural networks.
Due to the way chain rule and back propagation work, gradients of multiple layers are multiplied together. If most of the gradients of the activation units approach zero, the resultant gradient will approach zero as well. This phenomenon of vanishing gradient will in effect update the weights of the network only by an extremely small amount and cause the model to learn very slowly.

3. Why are sigmoid-like activation functions more prone to gradient vanishing?
The derivatives of sigmoid-like activation functions are smaller than one and close to zero mostly everywhere. Hence, when the model is poorly initialized or when the learning rate is too big, the derivatives approach zero easily during training. As opposed to sigmoid-like activation functions, the ReLU family of activation functions has constant one when input is greater than zero.
4. Calculate the derivative of Sigmoid function.

$$f'(z) = f(z)(1 - f(z))$$

5. What are the properties that make swish activation function a successful activation unit?
(1) unbounded above (2) bounded below (3) non-monotonicity (4) differentiable everywhere

6. Assume that your parents are training a neural network and encountered with gradient vanishing. The training loss is very high. Your parents haven't taken this course. They are trying to use a 100-layer MLP to fit a 1-D function which consists of 10 segments. They use sigmoid as activation unit. The range of inputs is [50, 100]. And the network was initialized by 0. Give 3 possible solutions for your parents.
(1) use fewer layers; (2) change activation units; (3) use batch normalization; (4) Xavier/Kaiming initialization

7. Please describe the main principle in deriving Xavier/Kaiming initialization for neural networks' weights.
The variance of the inputs and outputs in one certain layer should stay unchanged / be in the "sweet spot" of the activation units