1 Definitions

1.1 Gradient descent

- gradient descent is an iterative optimization algorithm used to minimize a function
  by iteratively moving in the direction of *steepest descent* as defined by the *negative of the gradient*
- use gradient descent to update the parameters (weights w, b) of the model
- find the optimal weights that reduce the prediction error (minimize loss)

Gradient descent algorithm

- s1: initialize the weights with random values and calculate error
- s2: calculate the gradient (the change in error when the weights are changed by a very small amount);
- s3: adjust weights with their gradients; helps move the weights in the direction in which the error is minimized
- s4: use the new weights for prediction and to calculate the new error
- s5: repeat steps 2 to 4 untill no significant error reduction

1.2 Gradient ascent

The optimization alorithm that takes steps proportional to the *positive of the gradient*, thus approaching a local maximum of that function.

1.3 Overfitting

- when the model is trying too hard to capture the noise in the training dataset
- it models the training data too well
- it doesn't generalize well to new data
- **solutions** for overfitting (*high variance*)
  - get more data (data augmentation) (e.g. rotations, flipping, zooming, distortions in images)
  - try regularization
  - try a different network architecture
  - try early stopping

1.4 Underfitting

- the model fails to correctly model the training data
- it also doesn't generalize to new data
- **solutions** for underfitting (*high bias*)
  - use a more complex model or a deeper model
    - try bigger network
    - try a different network architecture
  - train it longer
  - try some optimization algorithms

1.5 Vanishing and exploding gradients

- when training very deep neural networks the derivatives can end up either very very big or very very small, which makes training difficult
- the derivatives might increase exponentially or decrease exponentially as a function of L (number of layers), depending on the wights initial values
- make very carefull choices when initializing the weights in order to significantly reduce this problem