Loss functions

1. [KL Divergence](https://www.youtube.com/watch?v=SxGYPqCgJWM):
   1. Measures difference between two probability distributions P, Q
   2. Naïve approach: compare the likelihood to observe outputs generated from the different probabilities:
   3. Derivation:
      1. Let be the number of observations of event i,  
         and the respected probabilities of observations of event , over the distribution P, Q
      2. Now we can normalize this expression to the number of observations :
      3. And also take the log of this expression (possibly for numeric stability of probability values?)
      4. If we assume the number of observations goes to infinity, we get:  
          . Note that in this assumption we refer to P as the “true” distribution and measure how different Q is from P. (isn’t it just the likelihood definition? Why do we need this limit assumption?), So:
2. [Cross-entropy](https://www.youtube.com/watch?v=Pwgpl9mKars):
   1. A classification loss function
   2. Naïve approach: Assume model with parameter , input data and class y, minimize the difference between the predicted class distribution and the true class distribution . We can use the KL Divergence to measure this distance, and minimize it.
   3. Derivation:
      1. The KL divergence of this classification problem, where p is the true class distribution and q is the predicted class distribution:
      2. The first term doesn’t depend on so to minimize the entire function we can only minimize the second term, which is the cross entropy loss
      3. Cross entropy using sigmoid: we calculate the probabilities using the sigmoid function: . Therefore, the cross entropy will be:  
         where the last transition derived from p(y) being 0 for all class y that is not the true class, and 1 for the true class