**Adversarial Learning Course – Ex. 6 Part 2**

Zvika Deutsch, Elad Eatah

In this exercise, a white-box version of the Hop-Skip-Jump attack is implemented. This part presents an end-to-end implementation of the Hop-Skip-Jump attack. As in the previous part, the three main components are the images-interpolation stage, the binary-search stage (for finding the decision-boundary between two different classes) and the gradient-calculation stage

**General Description of the Attack**

Given two images which belong to two different labels of MNIST, this attack aims to generate adversarial examples for a given classifier. This attack relies on the existance of some boundary between images of the two different classes. This attack attempts to find images in this boundary, and these images are the output adversarial images.

The first step is the image-interpolation, where the hyper-line between the images is traversed. This line is guaranteed to intersect the 'decision-boundary'. Then, using binary-search, this algorithm seeks images whose classifications belong to neither of the classes are found. This search is stopped when such images are found, or when the length of the search-interval is smaller than some positive constant. This stage is implemented concurrently for multiple pairs of such images. Any images found in this stage are part of the output adversarial examples.

Then this white-box attack estimates the gradient of the function , defined in the previous part, valued in the found adversarial-images. This gradient is perpendicular to the decision-boundary. The algorithm then estiamates some step-size, and moves in the direction of this gradient. This new image is used as an endpoint for the next iteration for this algorithm, which may generate more adversarial-examples.

**Experimental Results**

In the attached notebook, 99 iterations of the attack on 100 randomlyy-chosen images from the MNIST test-set are presented (as a targeted attack). 95% of the images are moved-out successfully from their true-classes. The mean-radius of the perturbations of this attack is 0.25.