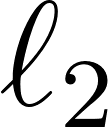
Assignment 6 - Part 2 - Hop Skip Jump

Eitamar Saraf, Moshe Nasletashvili

In this exercise we have dealt with the Hop Skip Jump attack. Before getting into technicalities, we would like first to explain the connection between decision boundaries and adversarial attacks.

The decision boundary is the surface on which classification is changed between one prediction to another. In our quest to find adversarial examples, we would like to find the closest (in some sense) example to the source that changes the model prediction. In this method, we search for the boundary point which is the closest (in the same vague sense) to the source. By adding a small perturbation to this boundary point, we could pass to the “other side” of the decision boundary and get to a close enough example that is classified differently by our model.

In our work, instead of tuning hyperparameters we have decided to take the paper (“HopSkipJumpAttack: A Query-EfficientDecision-Based Attack”, Chen et. al) method and change two points in the “BoundarySlide” function inspired by the paper:

1. The variable `perpendicular\_step` is now set to be iteration dependent, and decreases exponentially with the number of iterations.
2. We normalized the estimation of the decision boundary gradient by its [](https://www.codecogs.com/eqnedit.php?latex=%20%5Cell_2#0) norm - in this way we have a complete control of the step size we take.
3. In addition, we tuned the initialization of `perpendicular\_step` to be 0.1 instead of 0.05

Our results are as follows:

Test Attack report:

Source index: 1751 Source class: 4 Target class: 2

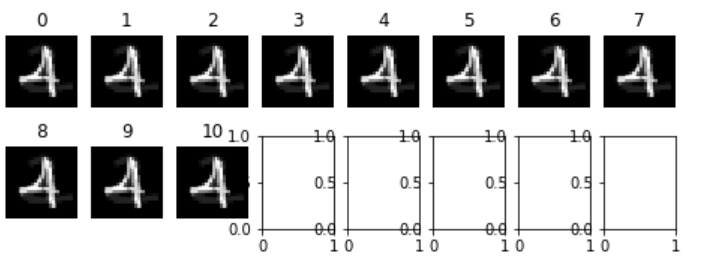
Mean boundary distance: 9.91

Successfully moved out of source class: 1.00

Successfully perturbed to target class: 0.96

Mean perturbation distance: 0.07

We can see that the mean boundary distance is much bigger than the mean perturbation distance (2 orders of magnitude) - this means that on average our adversarial examples were significantly closer to the source than to the destination, as wanted.

Visualization of the boundary points along the iterative process:

We can see visually what we have seen in the numerics - we are much closer to the source label (4) than to the target label (2).

We can observe as well, that all examples were successfully attacked on the non targeted manner, and only 4 images were not classified as the target label.