**FGSM**

We can think of FGSM as the opposite of gradient descent where we try to reduce the value of the parameters if the slope is positive – pulling it to the local minima and increase the value of the parameters if the slope is negative-pushing it to the local minima.

Mathematically,

The workflow of the algorithm is as follow:

1) Get the gradient of the image

2) Randomly initialize the value of perturbation

3) Add the value of the gradient in the direction of the perturbation.

**Basic Iterative Method (BIM)**

Is the iterative method of FGSM where the perturbations are applied iteratively . After each iteration, the perturbed image is clipped to be within limit of original image so that perturbed image is in the neighborhood of the original image.

The workflow of the algorithm is as follow:

1) initialise perturbation to 0

2) The perturbed image will then be the same as the original image

3) Take the gradient of the perturbed image with respect to perturbation from the previous iteration and follow the operations as stated in the equation above

4) Apply to the perturbed image.

5) add the perturbation to the perturbed image (from previous iteration)

6)Repeat 3-5 until you reach a maximum number of iterations.

**Projected Gradient Descent (PGD)**

PGD is an extension of Basic Iterative Method, the difference is that after each iteration, the perturbation can be constraint my any norm whereas for Basic iterative method , it is constraint to The equation of PGD therefore is :

The workflow of the algorithm is as follow:

1) initialise perturbation to 0

2) The perturbed image will then be the same as the original image

3) Take the gradient of the perturbed image with respect to perturbation from the previous iteration and follow the operations as stated in the equation above

4) Project the perturbation to the constraint values

5) add the perturbation to the perturbed image (from previous iteration)

6)Repeat 3-5 until you reach a maximum number of iterations.

**Carlini & Wagner (C&W) attack**

C&W attack is a more sophisticated iterative attack that tries to both constraint the perturbed image with an norm as well as minimize the cost function. The problem is defined as

To simply put it , we try to minimize the distance between the perturbed image and the original image while ensuring that the classifier misclassifies the given image. D can follow any norm (0,1,2 or ). The difference between this approach and approaches discussed earlier is in the objective function. C&W uses a softplus objective function. Therefore, the final minimization problem becomes

**Momentum Iterative Method**

This approach introduces momentum to the equation of FGSM. The momentum is calculated as the accumulation of gradients for . That is

Where is the momentum factor , a hyperparameter.

The perturbated image is then ,

This concept has the same logic of gradient descent with momentum – faster convergence and escape from local minima – but in the opposite direction.

The following flow chart gives a summary of the above discussed attacks.

