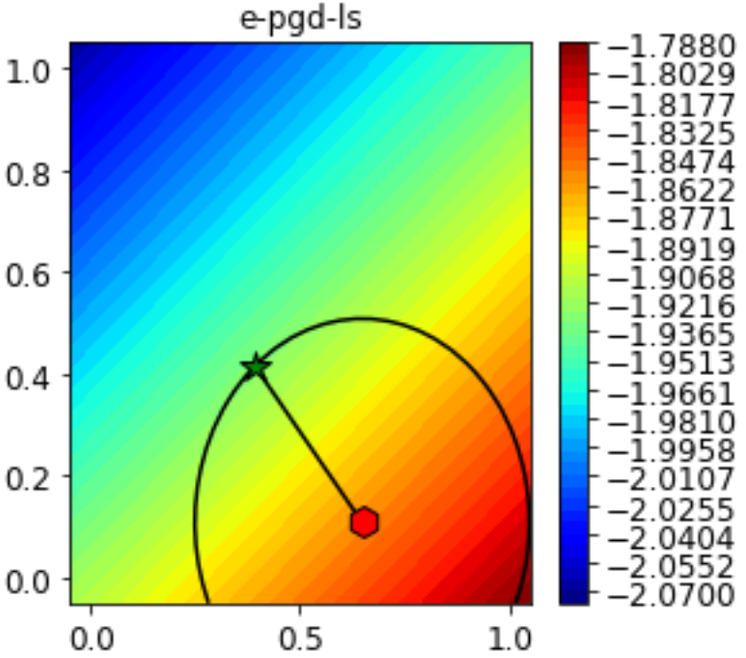
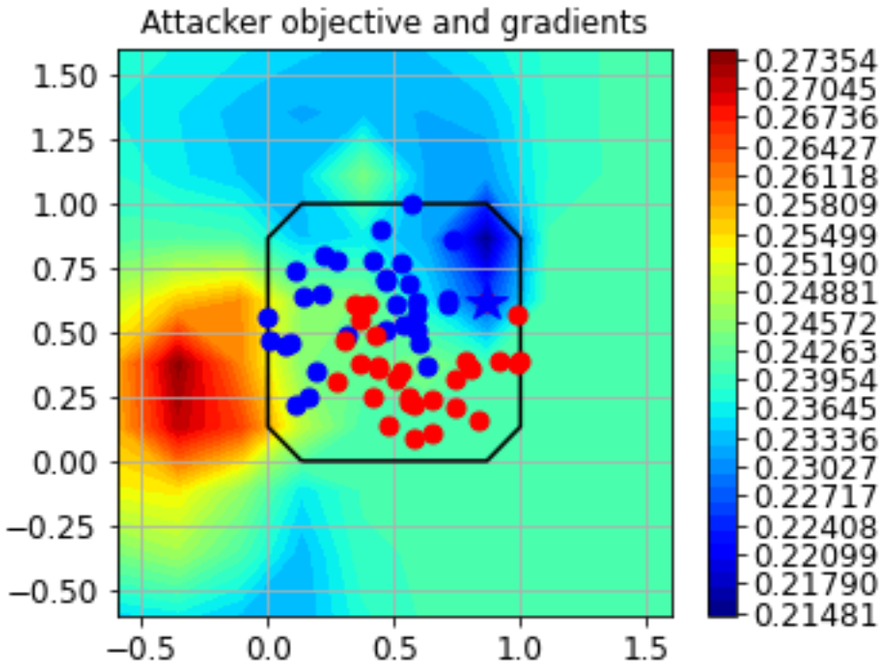
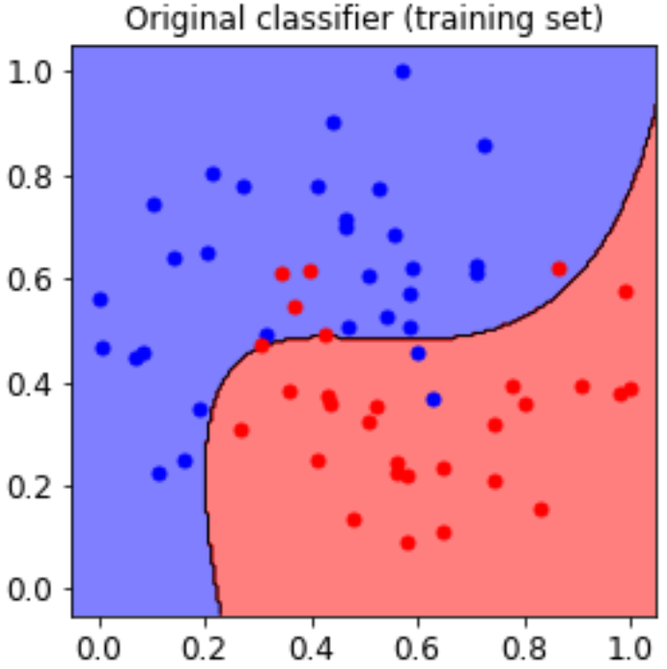
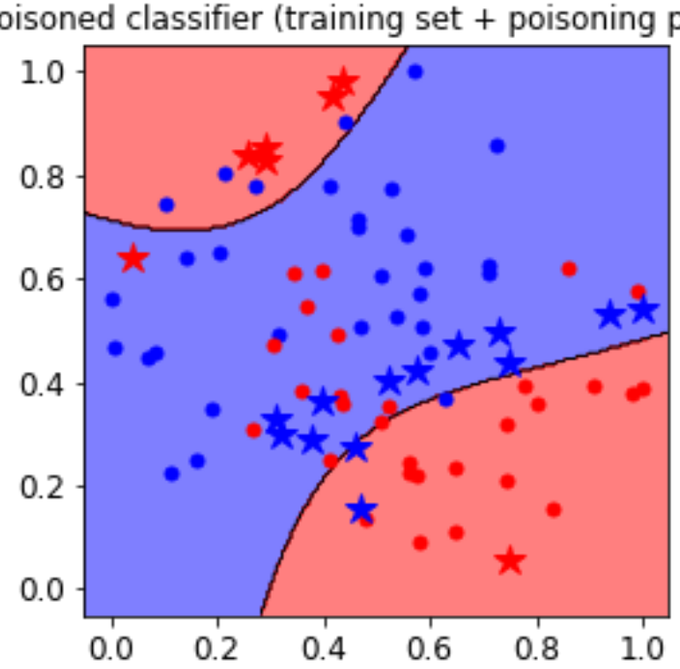

e-pgd-ls
−1.7880
−1.8029
−1.8177
−1.8325
−1.8474
−1.8622
−1.8771
−1.8919
−1.9068
−1.9216
−1.9365
−1.9513
−1.9661
−1.9810
−1.9958
−2.0107
−2.0255
−2.0404
−2.0552
−2.0700

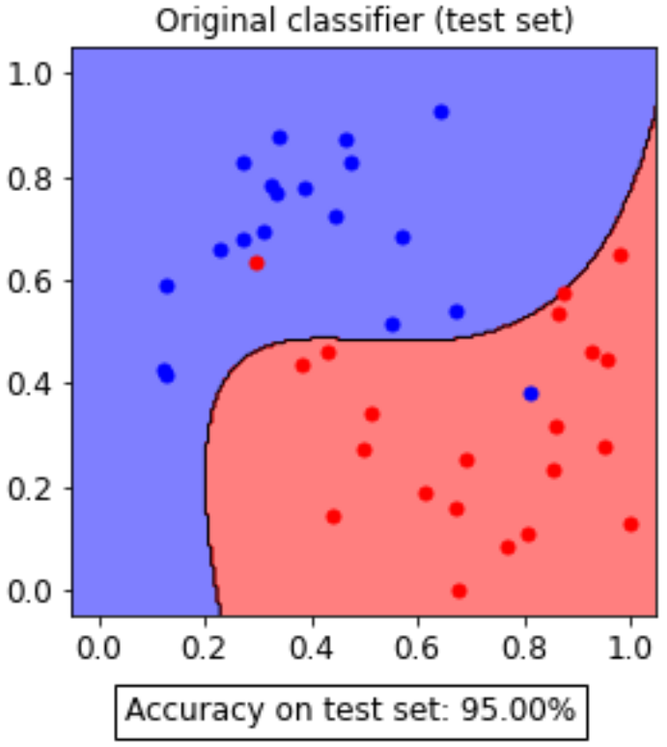
(a)

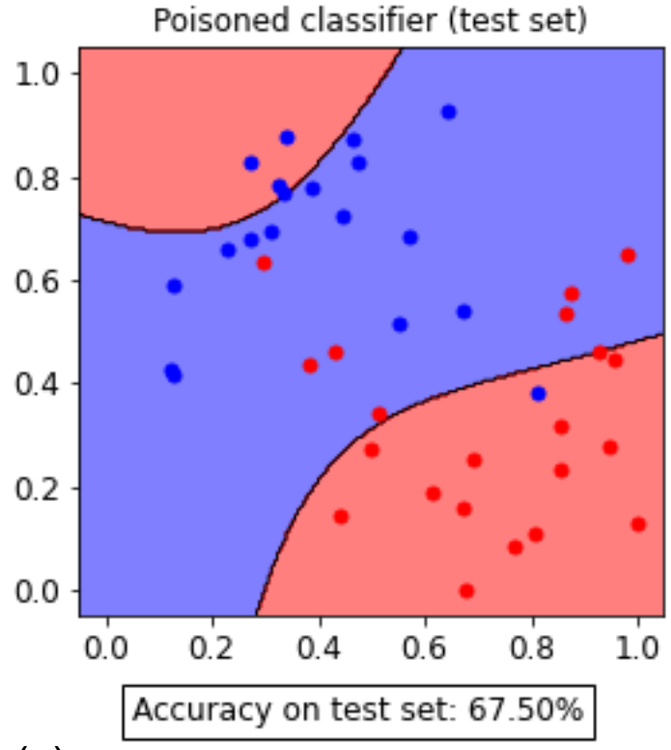

Attacker objective and gradients
0.27354
0.27045
0.26736
0.26427
0.26118
0.25809
0.25499
0.25190
0.24881
0.24572
0.24263
0.23954
0.23645
0.23336
0.23027
0.22717
0.22408
0.22099
0.21790
0.21481

(b)


Original classifier (training set)


Poisoned classifier (training set + poisoning points)


Original classifier (test set)
Accuracy on test set: 95.00%


Poisoned classifier (test set)
Accuracy on test set: 67.50%

(c)