Adversarial example with SVM support vector

Pranpaveen L.

01205519 Pattern Recognition

Kasetsart University

Adversarial Example

classified as

Stop Sign

+

X 0.07

=

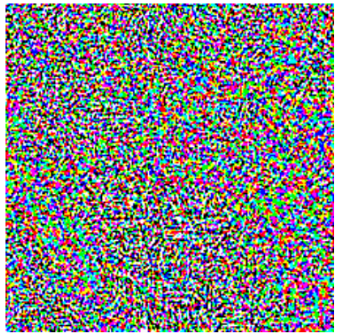
classified as

Max Speed 100

$\boldsymbol{x}$

"panda"
57.7% confidence

$+ .007 \times$



$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"nematode"
8.2% confidence

$=$

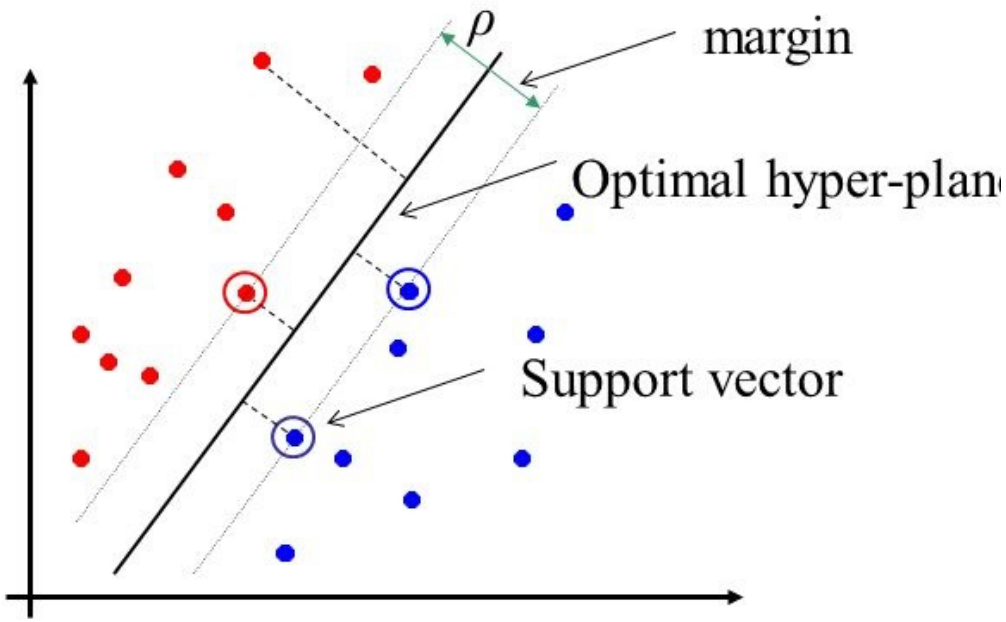
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$

"gibbon"
99.3 % confidence

https://deep.ghost.io/robust-attribution/

Hypotesis

- Support vector data in SVM is easier to attack



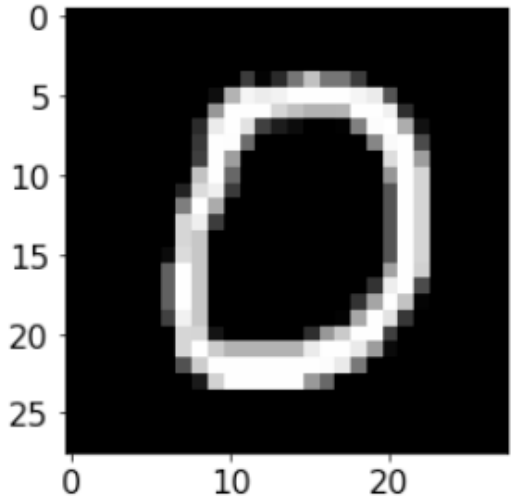
https://www.mitosistech.com/support-vector-machine/

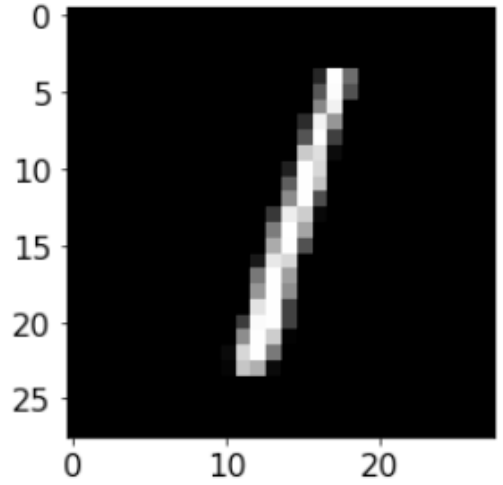
MNIST 2 class binary classification

```
mnistmodel_A2(
  (conv1): Conv2d(1, 64, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
  (conv2): Conv2d(64, 64, kernel_size=(5, 5), stride=(2, 2))
  (dense1): Linear(in_features=9216, out_features=32, bias=True)
  (dense2): Linear(in_features=32, out_features=2, bias=True)
  (dense3): Linear(in_features=2, out_features=1, bias=True)
)
```

```
----------------------------------------------------------------
      Layer (type)               Output Shape         Param #
================================================================
          Conv2d-1        [-1, 64, 28, 28]           1,664
          Conv2d-2        [-1, 64, 12, 12]         102,464
          Linear-3              [-1, 32]           294,944
          Linear-4               [-1, 2]                66
          Linear-5               [-1, 1]                 3
================================================================
Total params: 399,141
Trainable params: 399,141
Non-trainable params: 0
----------------------------------------------------------------
```

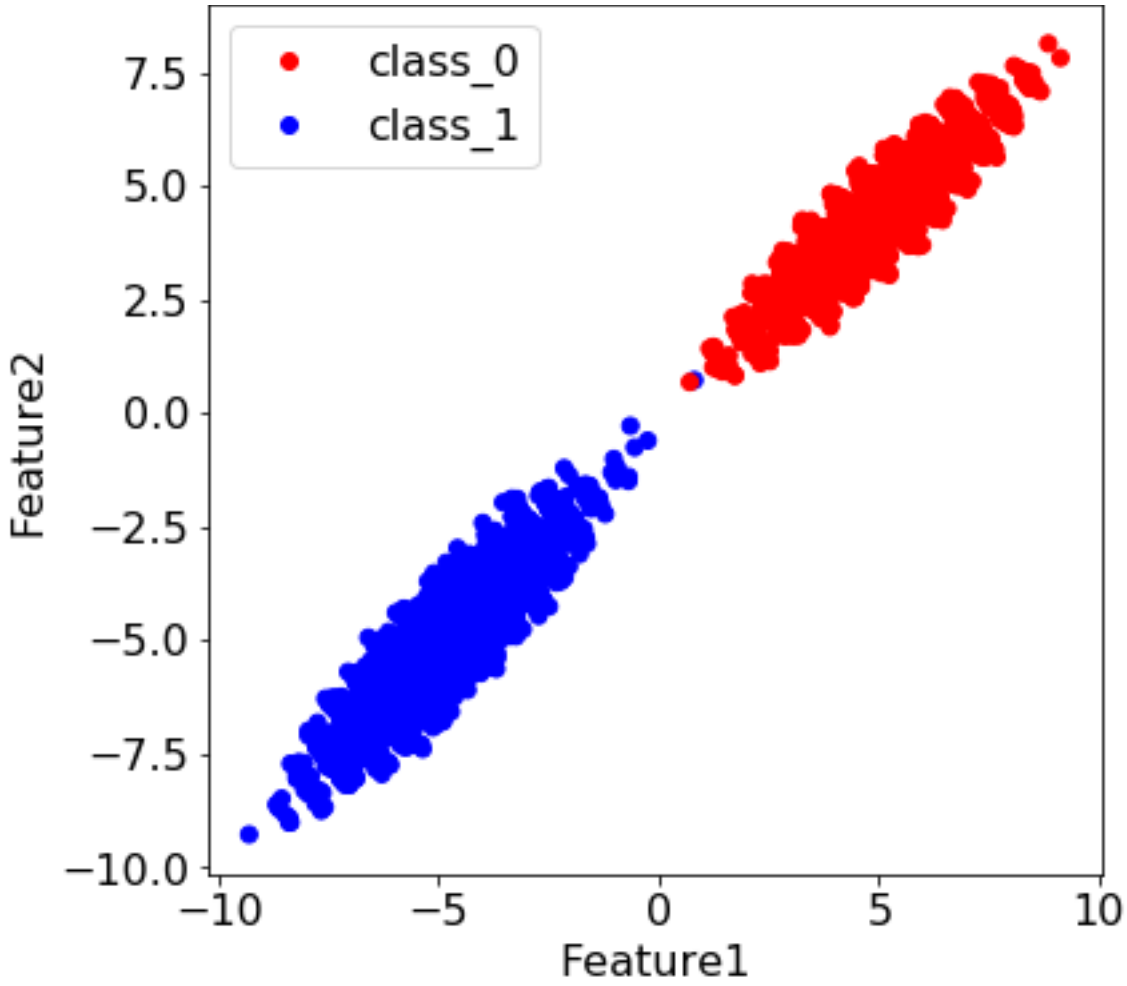
Feature extractor





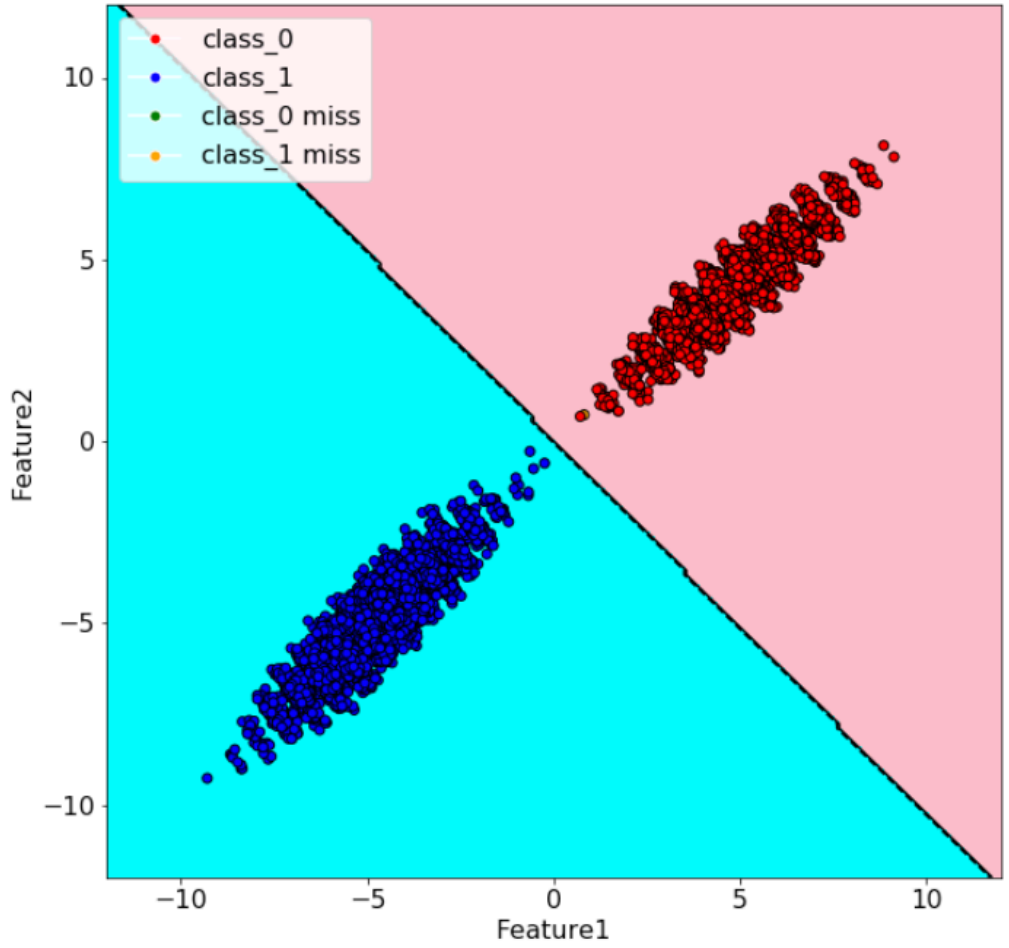
MNIST 2 class binary classification

Feature Extractor



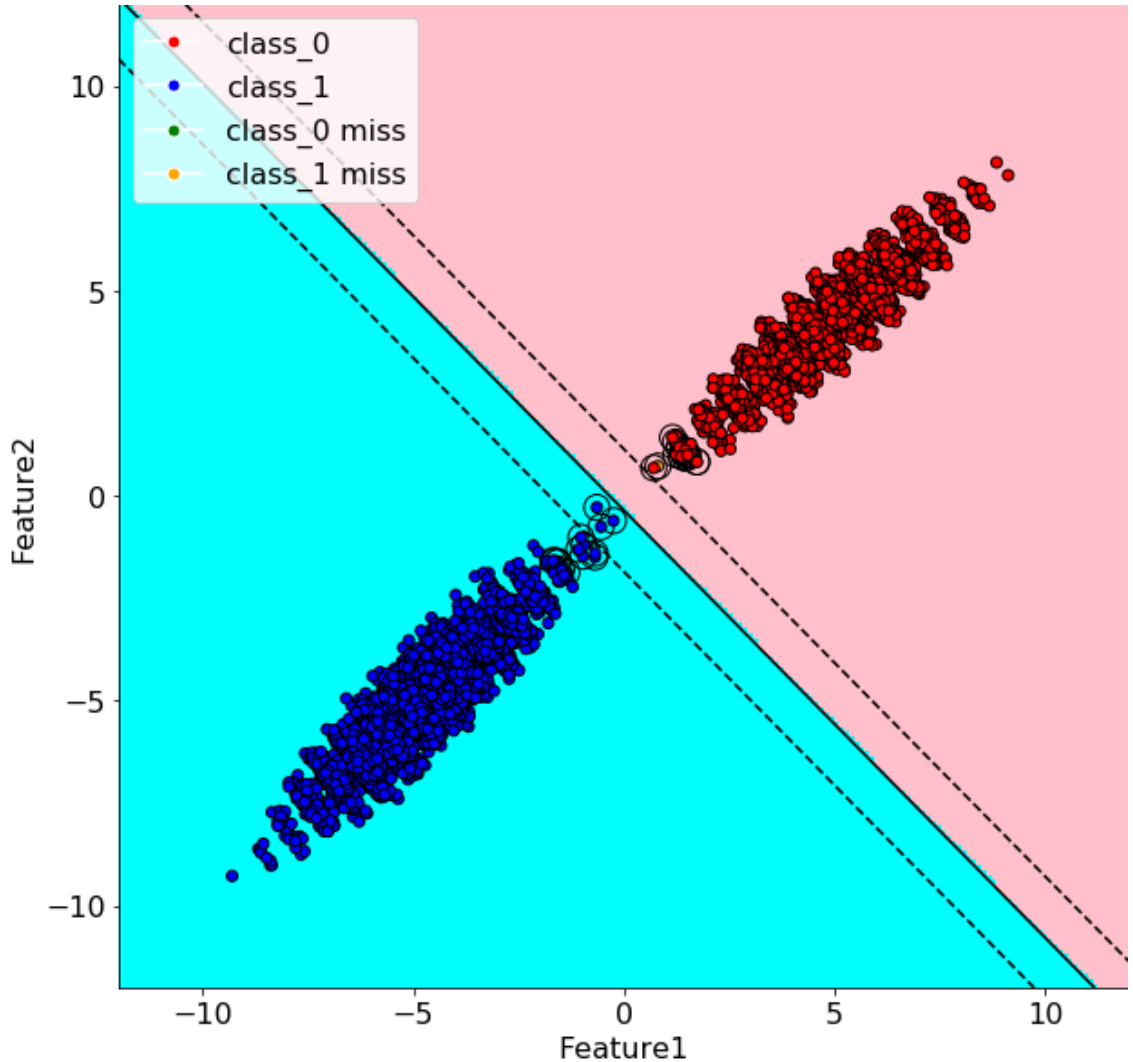
MNIST 2 class binary classification

- Last dnn layer with sigmoid

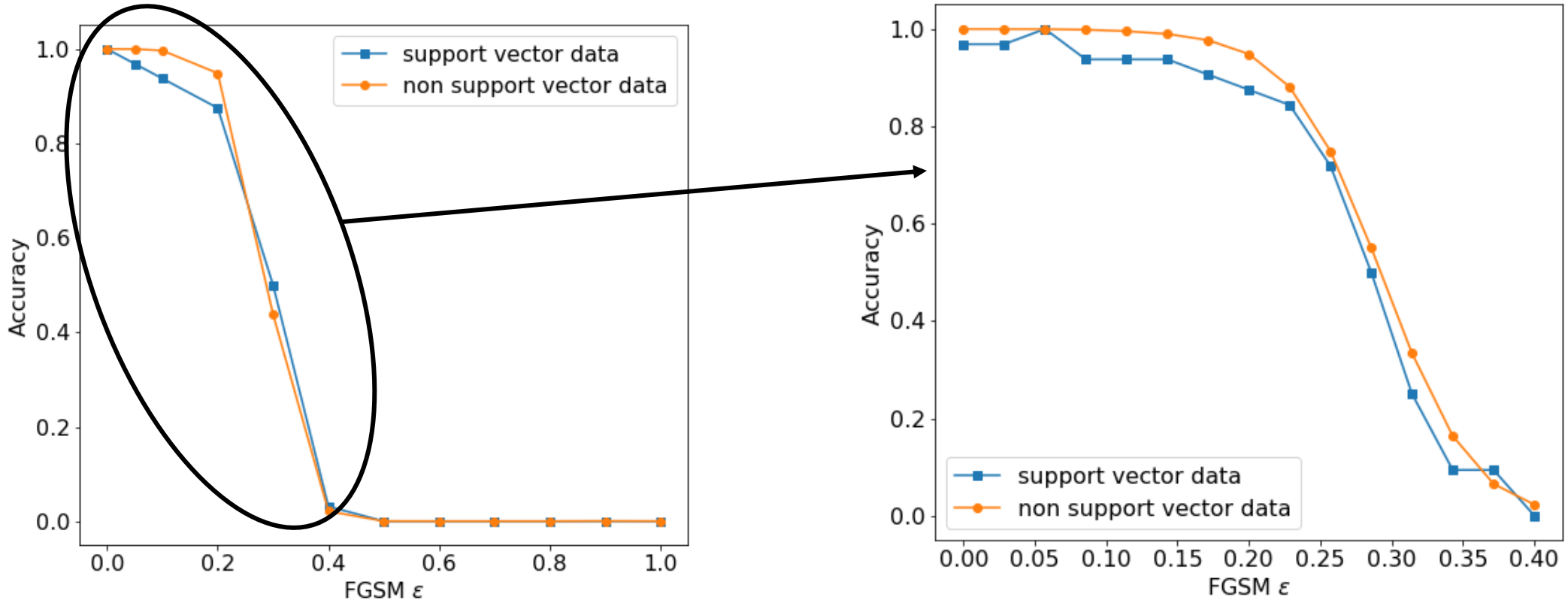


MNIST 2 class binary classification

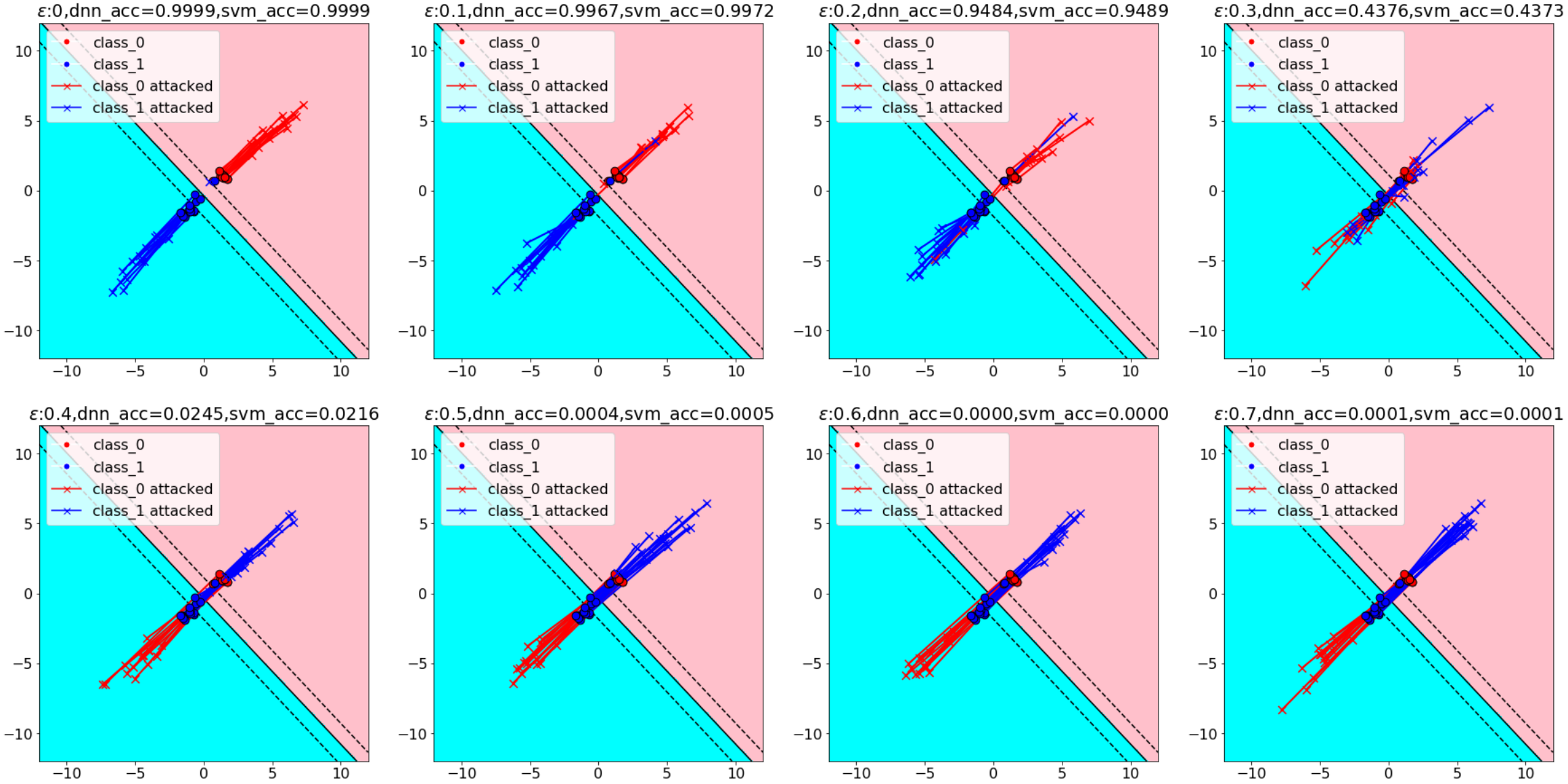
- SVM



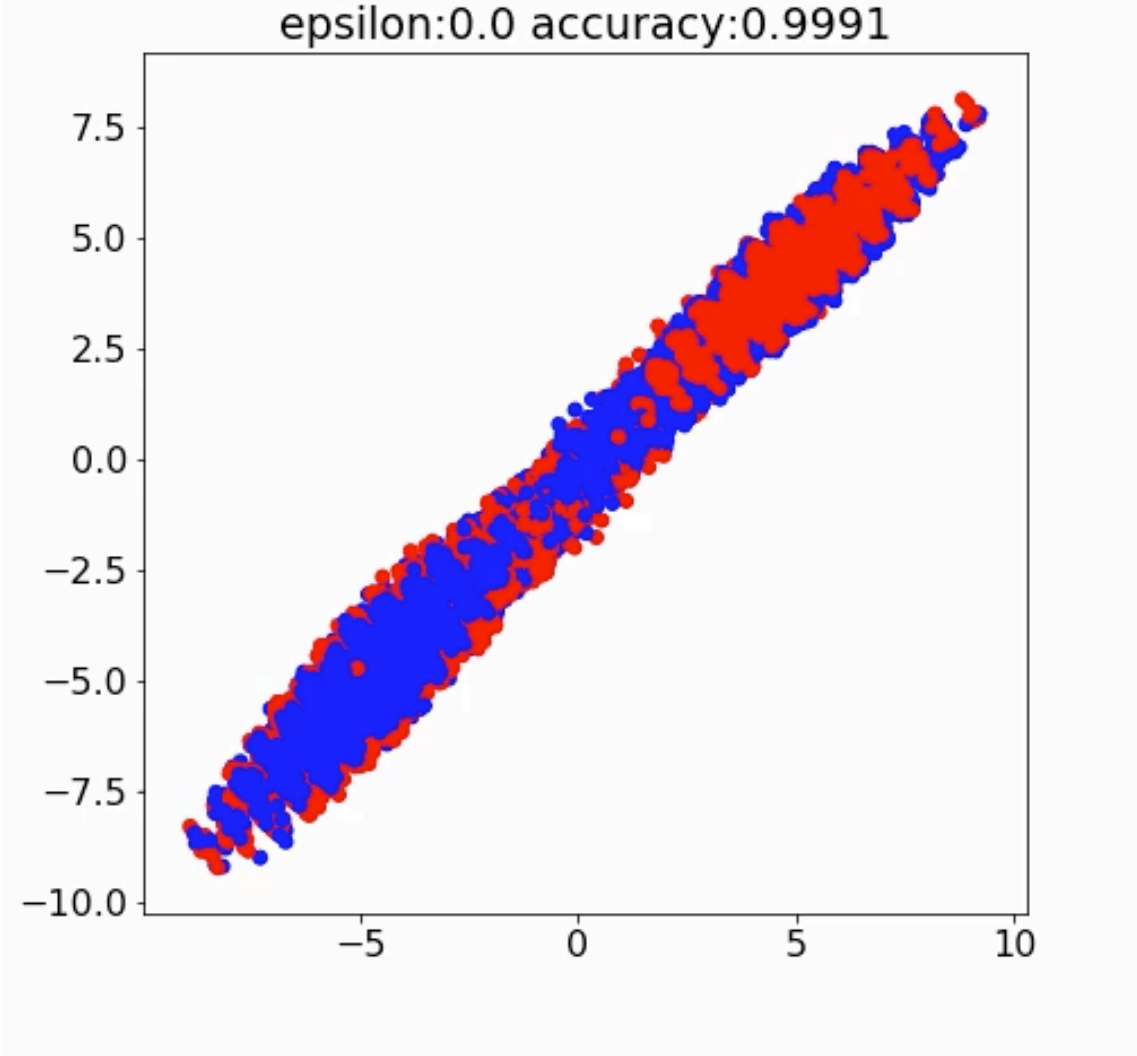
MNIST 2 class binary classification



MNIST 2 class binary classification



MNIST 2 class binary classification



MNIST dataset (10 class)

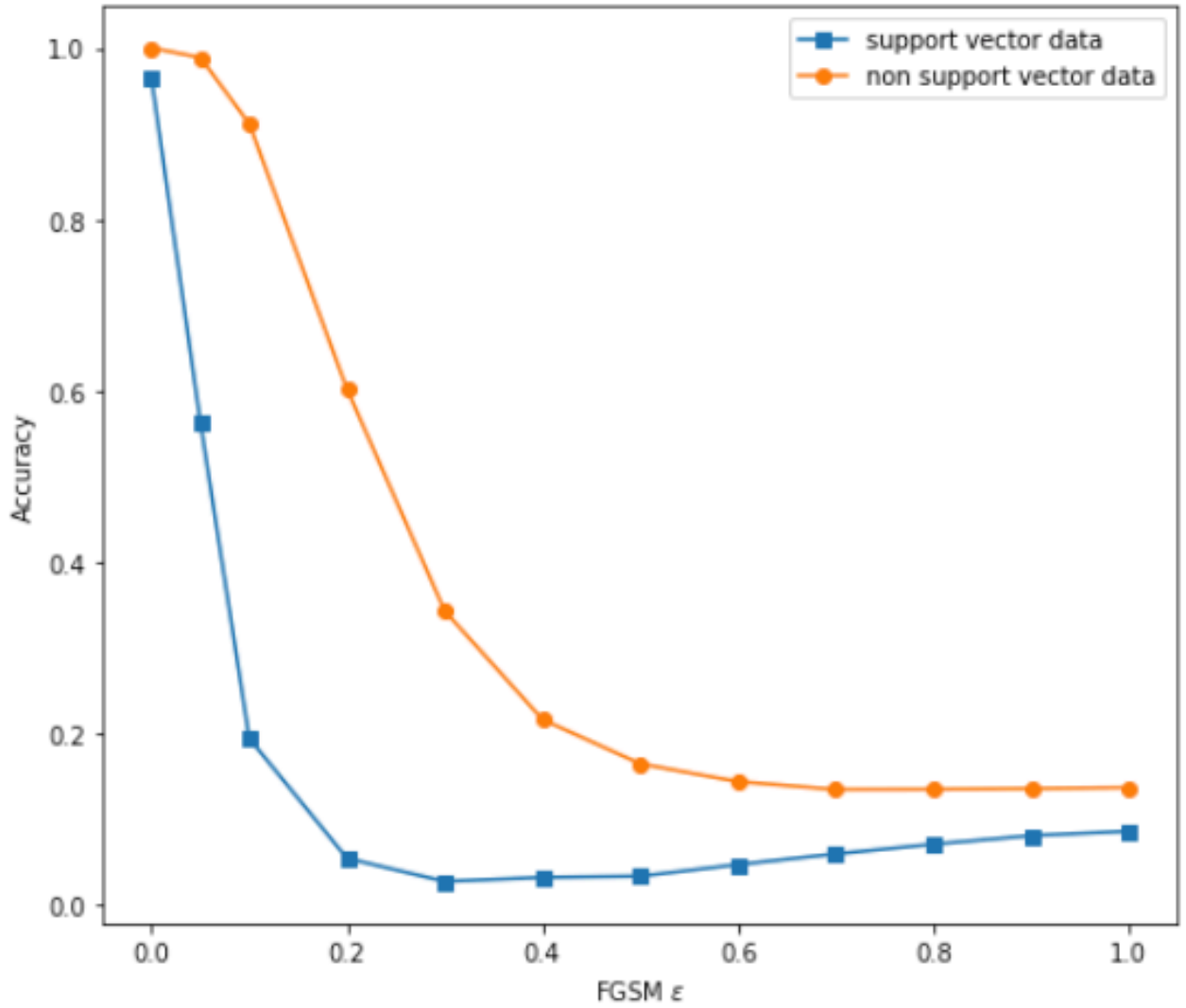
```
mnistmodel_A(
  (conv1): Conv2d(1, 64, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
  (conv2): Conv2d(64, 64, kernel_size=(5, 5), stride=(2, 2))
  (dense1): Linear(in_features=9216, out_features=128, bias=True)
  (dense2): Linear(in_features=128, out_features=10, bias=True)
)
```

Feature extractor

SVM (rbf kernel)

```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1          [-1, 64, 28, 28]           1,664
            Conv2d-2          [-1, 64, 12, 12]         102,464
            Linear-3                [-1, 128]       1,179,776
            Linear-4                 [-1, 10]           1,290
================================================================
Total params: 1,285,194
Trainable params: 1,285,194
Non-trainable params: 0
```

MNIST dataset (10 class)



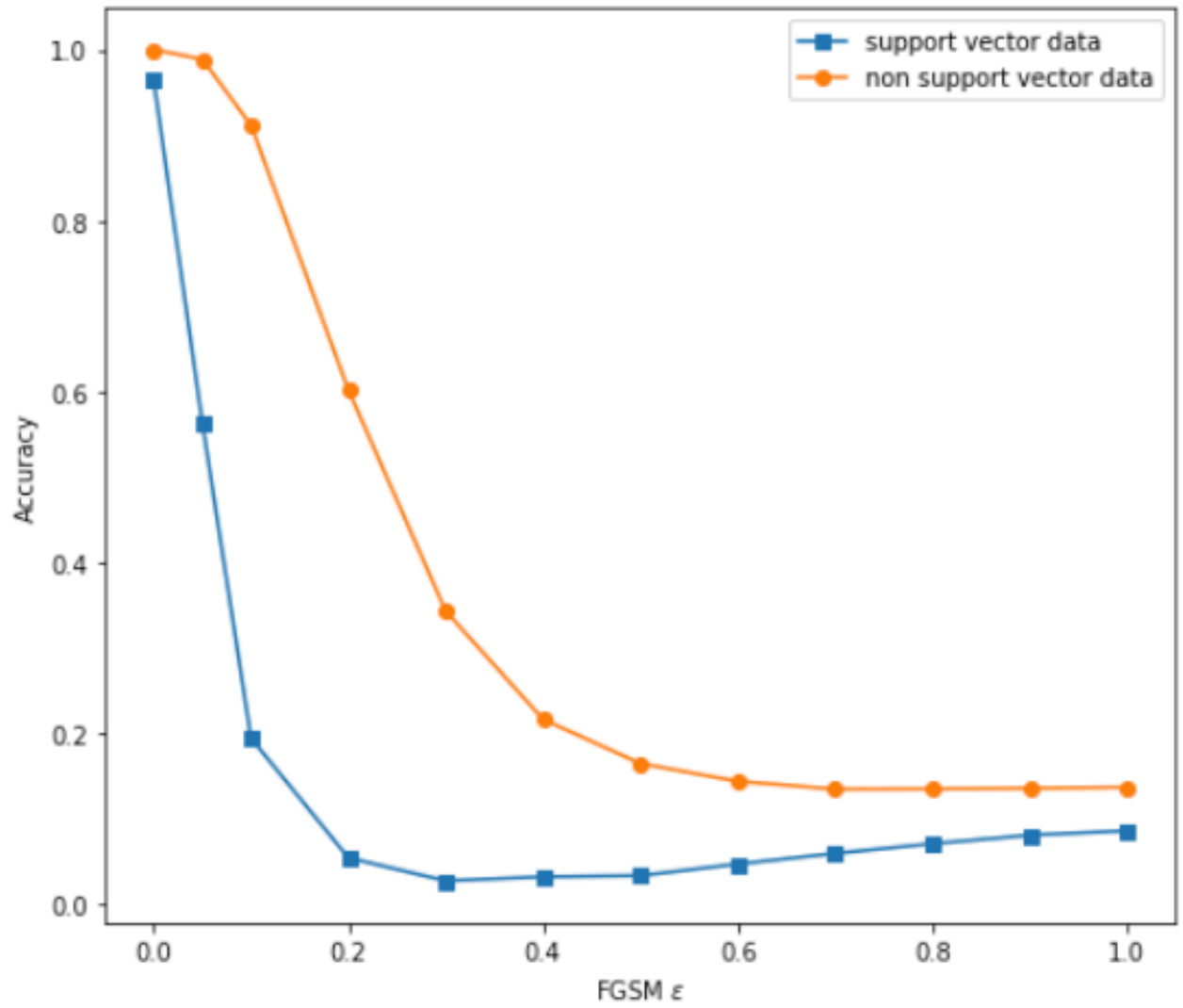
MNIST dataset (10 class)

```
mnistmodel_A(
  (conv1): Conv2d(1, 64, kernel_size=(5, 5), stride=(1, 1), padding=(2, 2))
  (conv2): Conv2d(64, 64, kernel_size=(5, 5), stride=(2, 2))
  (dense1): Linear(in_features=9216, out_features=128, bias=True)
  (dense2): Linear(in_features=128, out_features=10, bias=True)
)
```

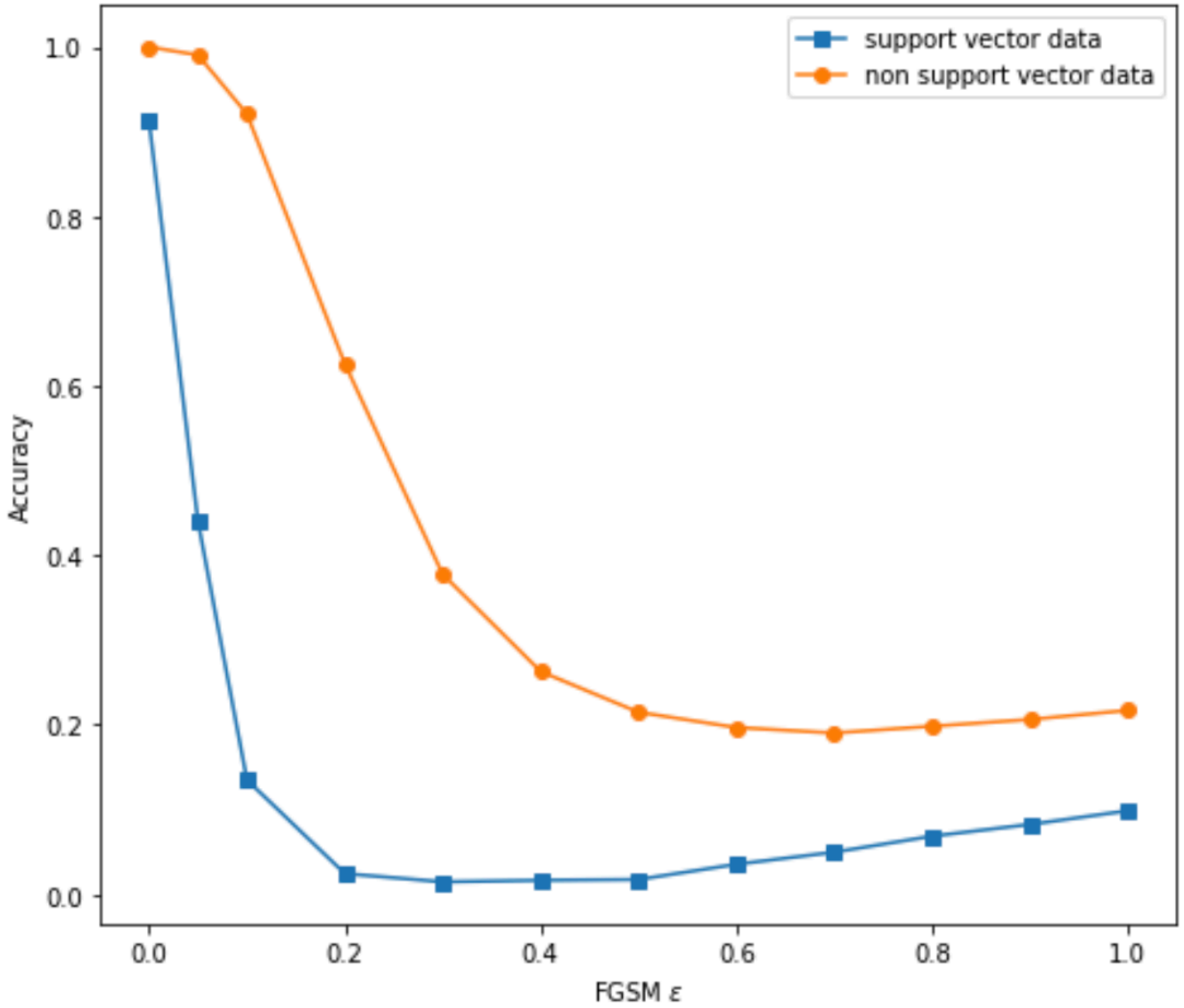
```
----------------------------------------------------------------
        Layer (type)               Output Shape         Param #
================================================================
            Conv2d-1           [-1, 64, 28, 28]           1,664
            Conv2d-2           [-1, 64, 12, 12]         102,464
            Linear-3                 [-1, 128]       1,179,776
            Linear-4                  [-1, 10]           1,290
================================================================
Total params: 1,285,194
Trainable params: 1,285,194
Non-trainable params: 0
```

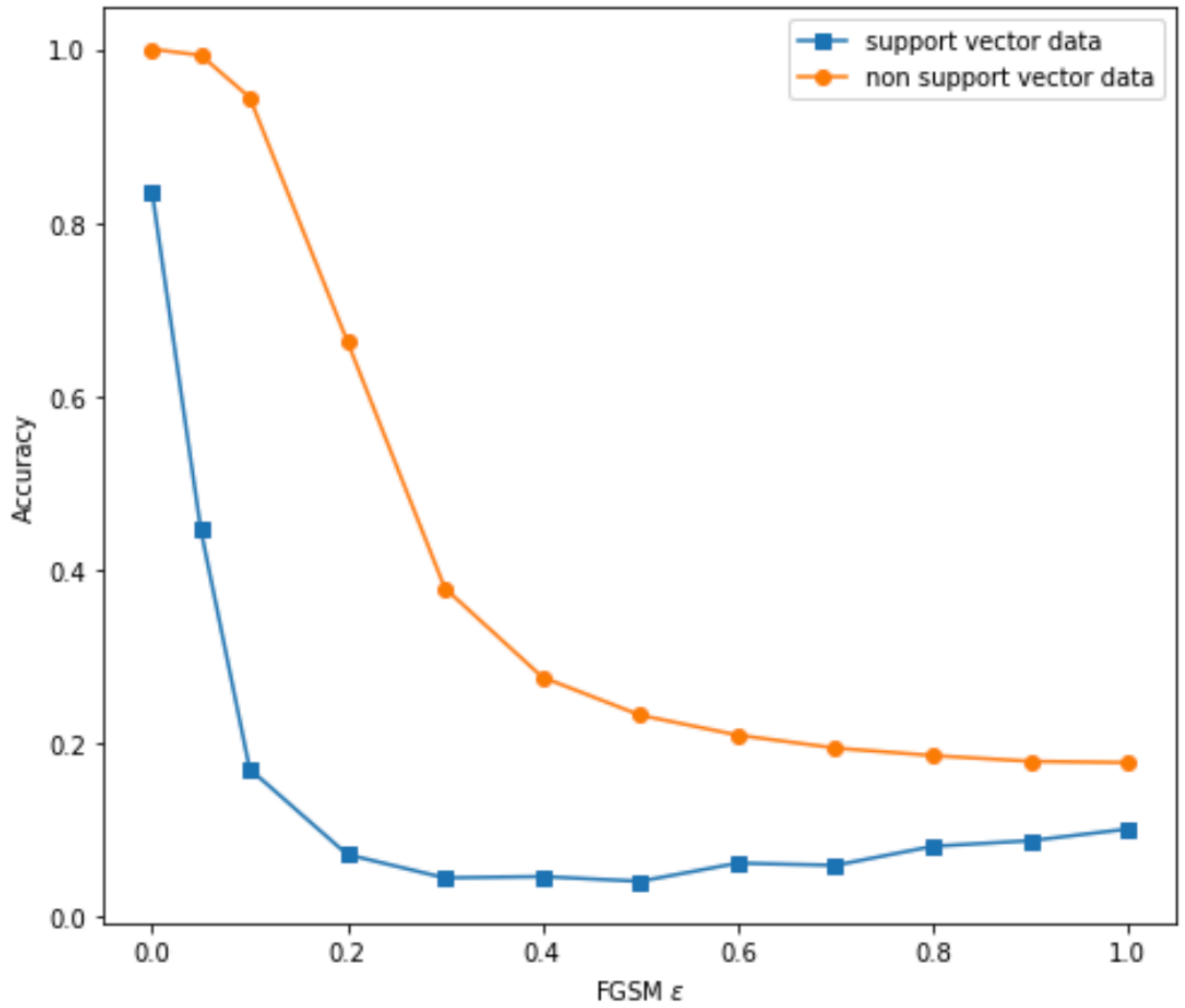
128
64
32
16
10

128

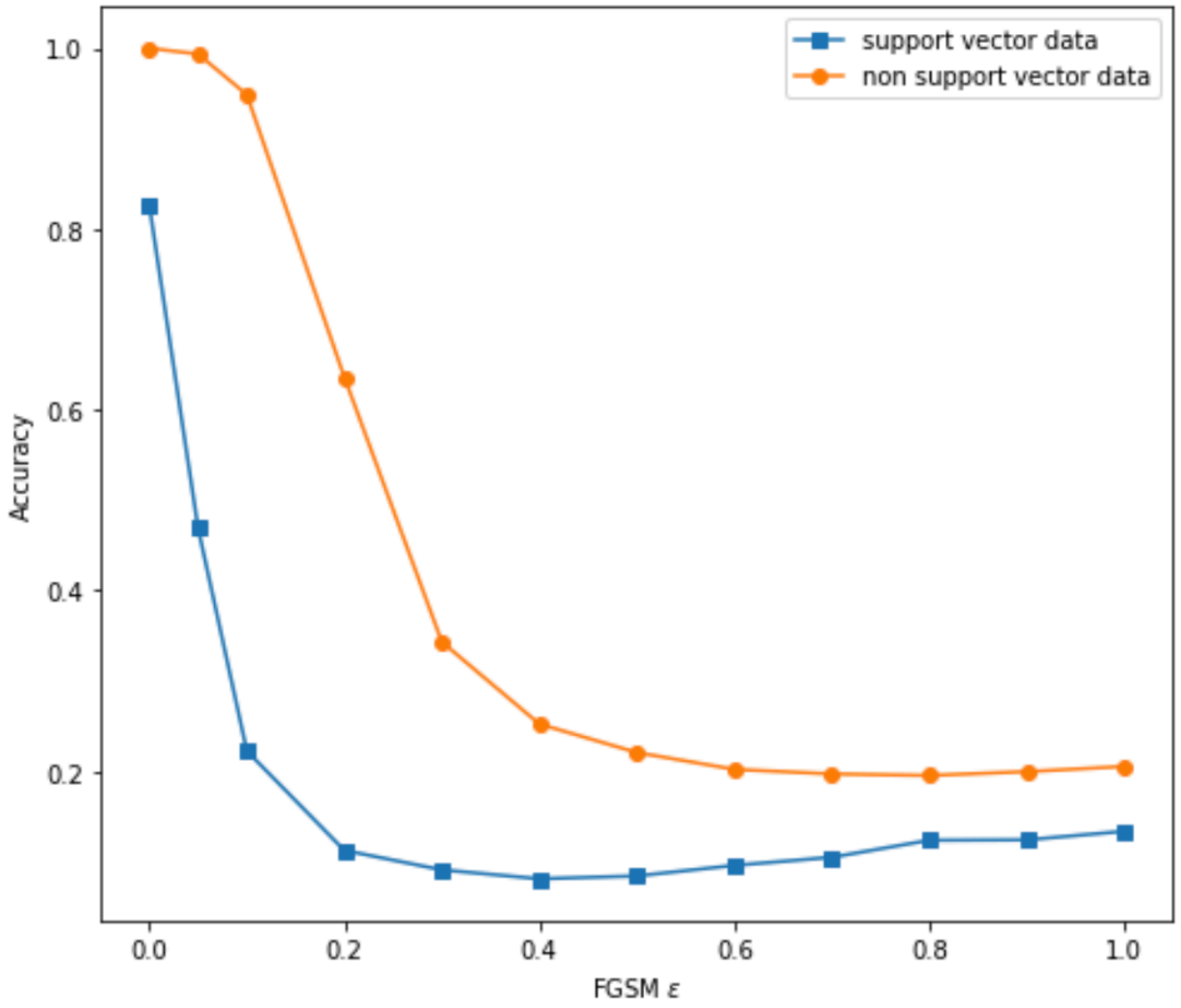

support vector data
non support vector data
Accuracy
FGSM ε
0.0
0.2
0.4
0.6
0.8
1.0

64

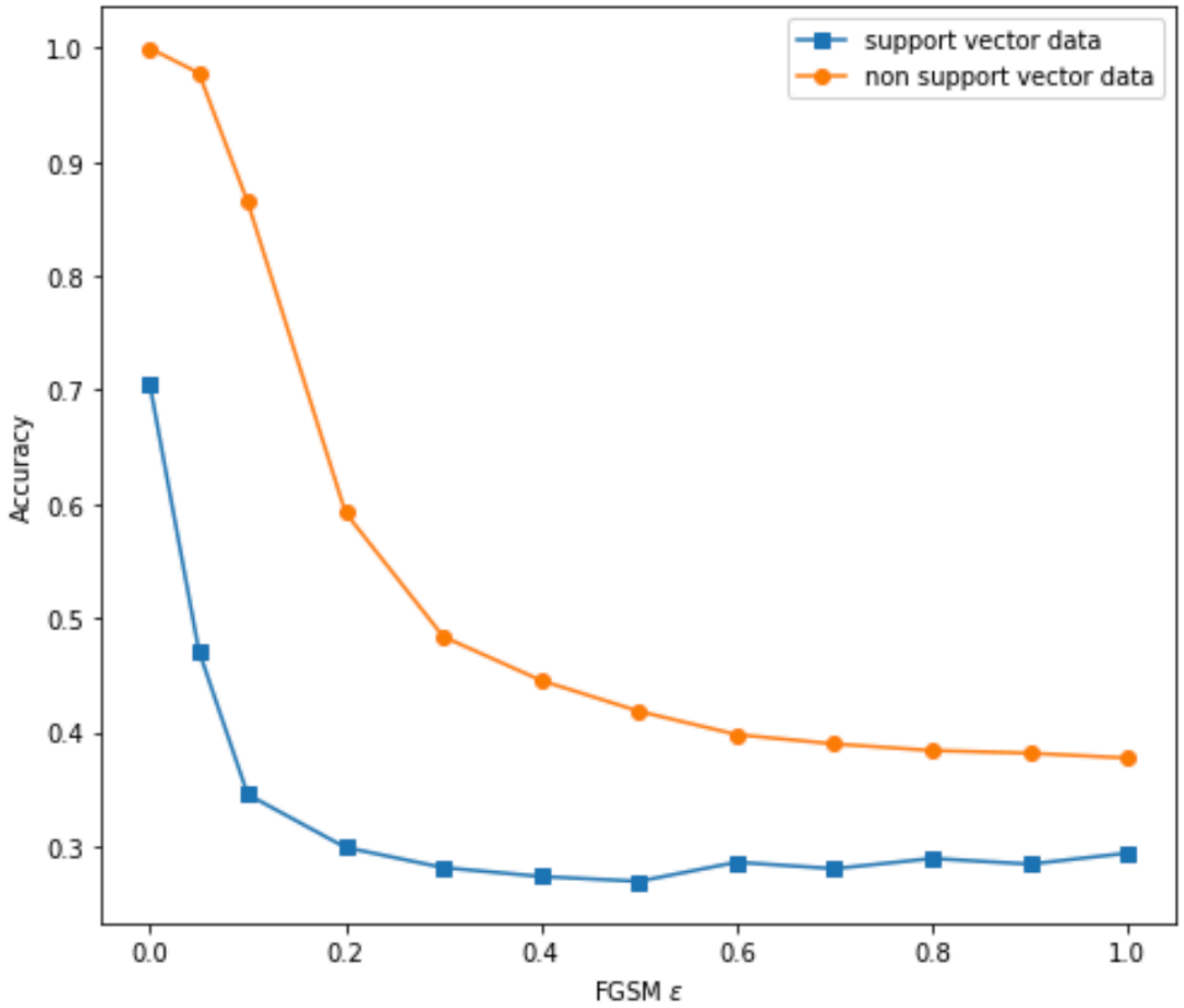

support vector data
non support vector data
Accuracy
FGSM ε
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


support vector data
non support vector data
Accuracy
FGSM ε
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

16


support vector data
non support vector data
Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.2
0.4
0.6
0.8
1.0
FGSM ε

10


support vector data
non support vector data
Accuracy
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.0
0.2
0.4
0.6
0.8
1.0
FGSM ε

Conclusion

Q/A