Dataset1-Regression_output_9

October 19, 2021

1 Dataset 1 - Regression

1.1 Experiment Details

The aim of the experiment is to verify if the: 1. ABC_GAN model corrects model misspecification 2. ABC_GAN model performs better and converges faster than a simple C-GAN model

In the experiment we predict the distribution that represents the real data and simulate realistic fake data points using statistical mode, C-GAN and ABC-GAN model with 3 priors. We analyze and compare their performance using metrics like mean squared error, mean absolute error , manhattan distance and euclidean distance between y_{real} and y_{pred}

The models are as follows:

1. The statistical model assumes the distribution $Y = \beta X + \mu$ where $\mu \sim N(0, 1)$
2. The Conditional GAN consists of
    1. Generator with 2 hidden layers with 100 nodes each and ReLu activation.
    2. Discriminator with 2 hidden layers with 25 and 50 nodes and ReLu activation. We use Adam's optimser and BCE Logit Loss to train the model. The input to the Generator of the GAN is (x,e) where x are the features and $e \sim N(0, 1)$. The discriminator output is linear.
3. The ABC GAN Model consists of
    1. ABC generator is defined as follows:
        1. $Y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + ... + \beta_n x_n + N(0, \sigma)$ where $\sigma = 0.1$
        2. $\beta_i \sim N(0, \sigma^*)$ when $\mu = 0$ else $\beta_i \sim N(\beta_i^*, \sigma^*)$ where $\beta_i^* s$ are coefficients obtained from statistical model
        3. σ^* takes the values 0.01,0.1 and 1

    2. C-GAN network is as defined above. However the input to the Generator of the GAN is (x, y_{abc}) where y_{abc} is the output of the ABC Generator.

1.2 Import Libraries

[1]:
```
import warnings
warnings.filterwarnings('ignore')
```

[2]:
```
import train_test
import ABC_train_test
import regressionDataset
import network
```

```
import statsModel
import performanceMetrics
import dataset
import sanityChecks
import torch
import matplotlib.pyplot as plt
import seaborn as sns
from scipy.stats import norm
from torch.utils.data import Dataset,DataLoader
from torch import nn
```

1.3 Parameters

General Parameters

1. Number of Samples
2. Number of features

ABC-Generator parameters are as mentioned below: 1. mean : 1 ($\beta \sim N(\beta^*, \sigma)$ where β^* are coefficients of statistical model) or 1 ($\beta \sim N(0, \sigma)$ 2. std : $\sigma = 1, 0.1, 0.01$ (standard deviation)

[3]:
```
n_features = 10
n_samples= 100

#ABC Generator Parameters
mean = 1
variance = 0.001
```

[4]:
```
# Parameters
n_samples = 100
n_features = 10
mean = 0
variance = 1
```

1.4 Dataset

Generate a random regression problem

$Y = 1 + \beta_1 x_1 + \beta_2 x_2 + \beta_2 x_3 + ... + \beta_n x_n + N(0, \sigma)$ where $\sigma = 0.1$

[5]:
```
X,Y = regressionDataset.regression_data(n_samples,n_features)
```

```
         X1        X2        X3        X4        X5        X6        X7  \
0 -0.300035 -0.950023 -1.552244  0.069784  0.490576  0.234620 -0.836066
1  1.511078  0.466665  0.472634 -1.323431  0.188805 -0.644583 -0.355095
2  1.507415  1.685880 -0.490751 -1.407887 -0.389036 -1.951134  1.394092
3  1.644784  0.128439 -2.483455 -0.325933 -2.127511 -0.657200 -0.218824
4 -1.742249 -1.134471 -0.174136  1.125129  1.035036 -0.905433  0.090559

         X8        X9       X10           Y
```

```
0 -0.870214  0.703681  0.518100 -242.203792
1  0.641636 -0.732015  1.342412  139.711766
2  0.854711 -2.063695 -0.335557   83.939540
3 -0.784128 -0.297275 -1.382000 -371.079981
4 -0.035847  0.469781  0.119179 -150.373343
```

1.5 Stats Model

[6]:
```
[coeff,y_pred] = statsModel.statsModel(X,Y)
```

```
No handles with labels found to put in legend.
```

OLS Regression Results

Dep. Variable:	Y	R-squared:	1.000
Model:	OLS	Adj. R-squared:	1.000
Method:	Least Squares	F-statistic:	3.833e+07
Date:	Tue, 19 Oct 2021	Prob (F-statistic):	1.23e-290
Time:	23:26:07	Log-Likelihood:	621.89
No. Observations:	100	AIC:	-1222.
Df Residuals:	89	BIC:	-1193.
Df Model:	10		
Covariance Type:	nonrobust		

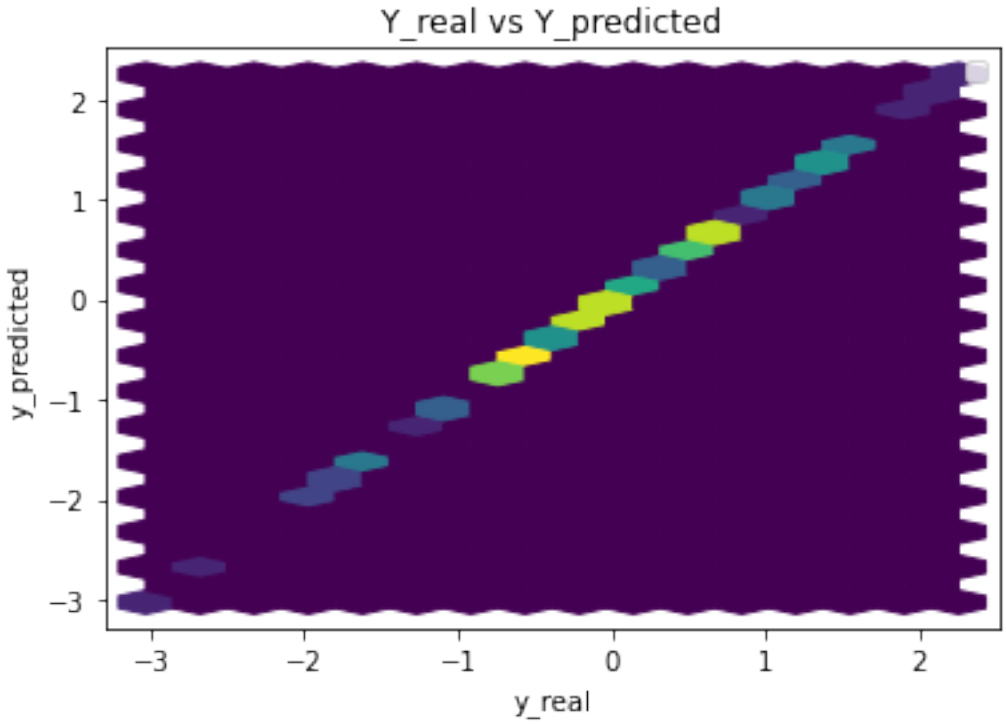
	coef	std err	t	P>|t|	[0.025	0.975]
const	9.021e-17	5.11e-05	1.77e-12	1.000	-0.000	0.000
x1	0.1248	5.54e-05	2252.712	0.000	0.125	0.125
x2	0.4942	5.35e-05	9240.436	0.000	0.494	0.494
x3	0.5776	5.25e-05	1.1e+04	0.000	0.578	0.578
x4	0.1132	5.3e-05	2136.443	0.000	0.113	0.113
x5	0.0763	5.64e-05	1353.177	0.000	0.076	0.076
x6	0.3331	5.21e-05	6386.981	0.000	0.333	0.333
x7	0.3384	5.18e-05	6534.653	0.000	0.338	0.338
x8	0.2230	5.48e-05	4067.697	0.000	0.223	0.223
x9	0.0227	5.64e-05	402.803	0.000	0.023	0.023
x10	0.4182	5.62e-05	7435.596	0.000	0.418	0.418

Omnibus:	0.302	Durbin-Watson:	2.196
Prob(Omnibus):	0.860	Jarque-Bera (JB):	0.147
Skew:	-0.093	Prob(JB):	0.929
Kurtosis:	3.023	Cond. No.	1.79

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

```
Parameters:  const    9.020562e-17
x1       1.248070e-01
```

```
x2        4.942317e-01
x3        5.776378e-01
x4        1.131731e-01
x5        7.627529e-02
x6        3.330697e-01
x7        3.383524e-01
x8        2.230452e-01
x9        2.272429e-02
x10       4.182417e-01
dtype: float64
```




```
Performance Metrics
Mean Squared Error: 2.3221604842124612e-07
Mean Absolute Error: 0.0003861832025725345
Manhattan distance: 0.03861832025725345
Euclidean distance: 0.004818880040229742
```

1.6 Common Training Parameters (GAN & ABC_GAN)

[7]:
```
n_epochs = 5000
error = 0.001
batch_size = n_samples//2
```

1.7 GAN Model

[8]:
```
real_dataset = dataset.CustomDataset(X,Y)
device = torch.device('cuda' if torch.cuda.is_available() else 'cpu')
```

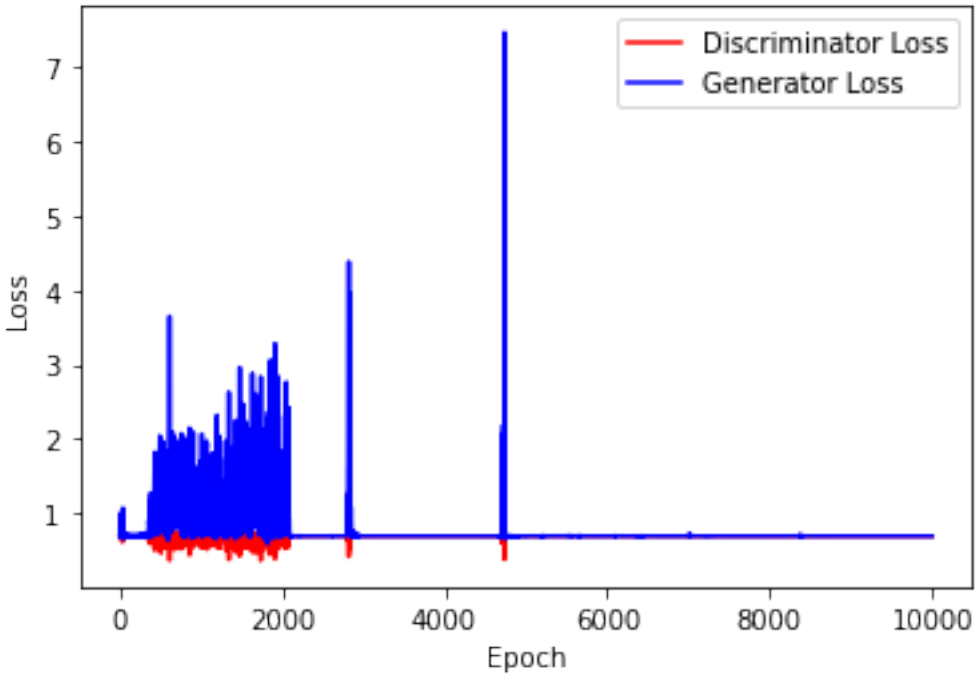
Training GAN for n_epochs number of epochs

[9]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
↪999))
```

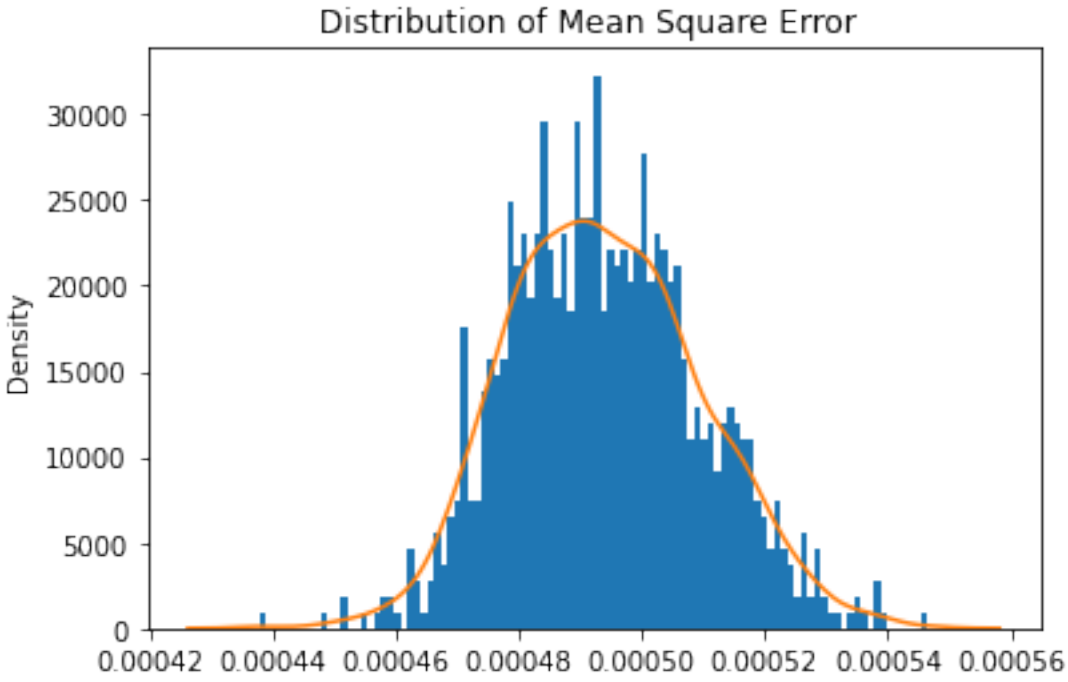
[10]:
```
print(generator)
print(discriminator)
```

```
Generator(
  (hidden1): Linear(in_features=12, out_features=100, bias=True)
  (hidden2): Linear(in_features=100, out_features=100, bias=True)
  (output): Linear(in_features=100, out_features=1, bias=True)
  (relu): ReLU()
)
Discriminator(
  (hidden1): Linear(in_features=12, out_features=25, bias=True)
  (hidden2): Linear(in_features=25, out_features=50, bias=True)
  (output): Linear(in_features=50, out_features=1, bias=True)
  (relu): ReLU()
)
```

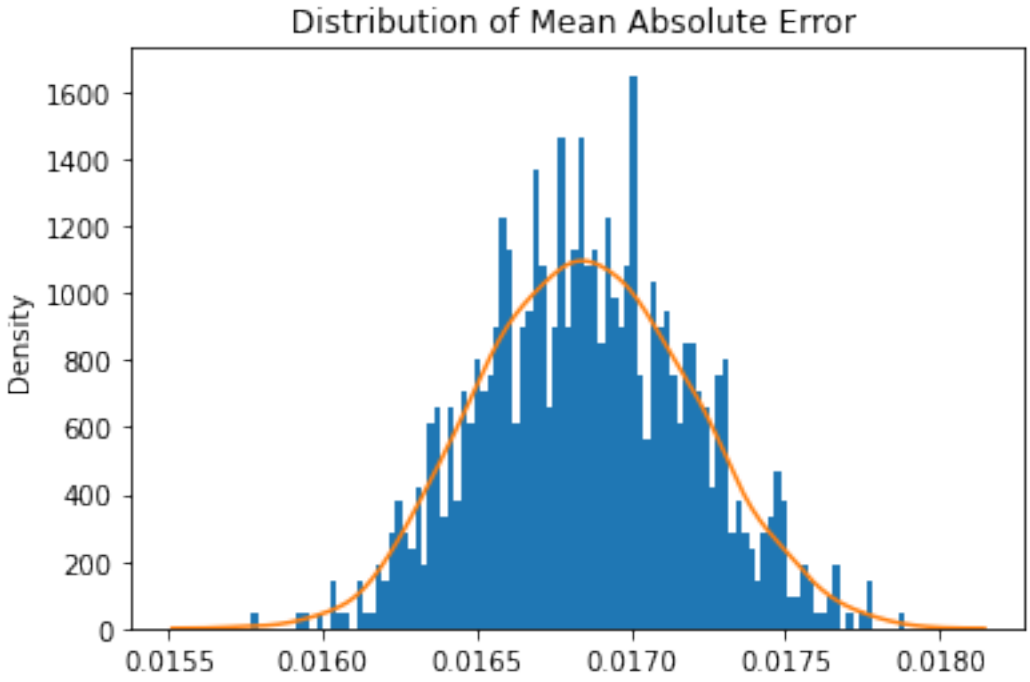
[11]:
```
train_test.
↪training_GAN(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,␣
↪n_epochs,criterion,device)
```



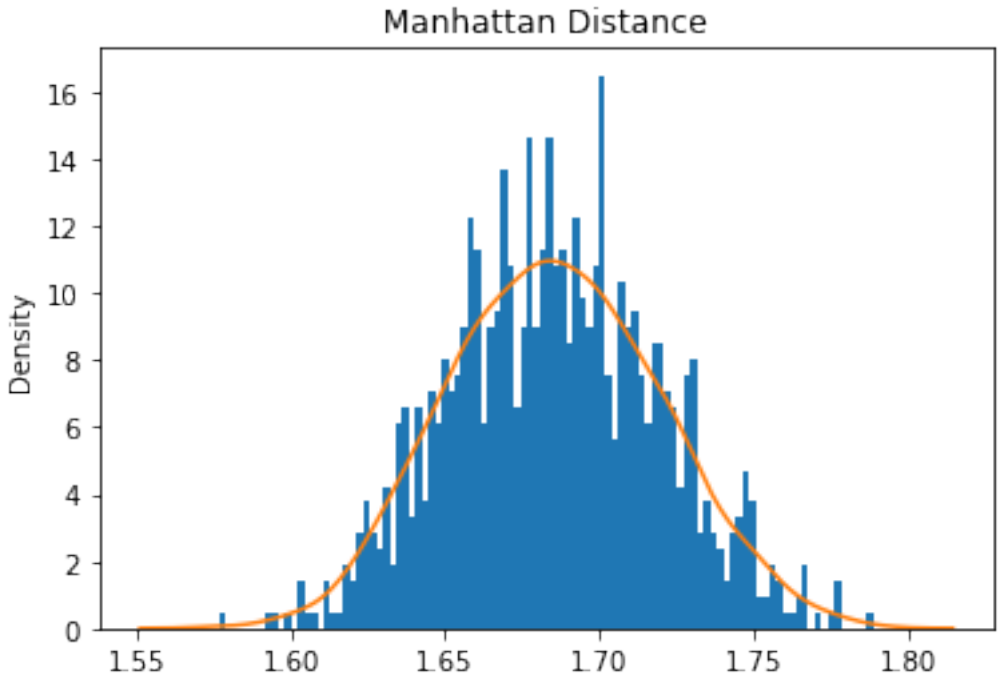
[12]:
```
train_test.test_generator(generator,real_dataset,device)
```



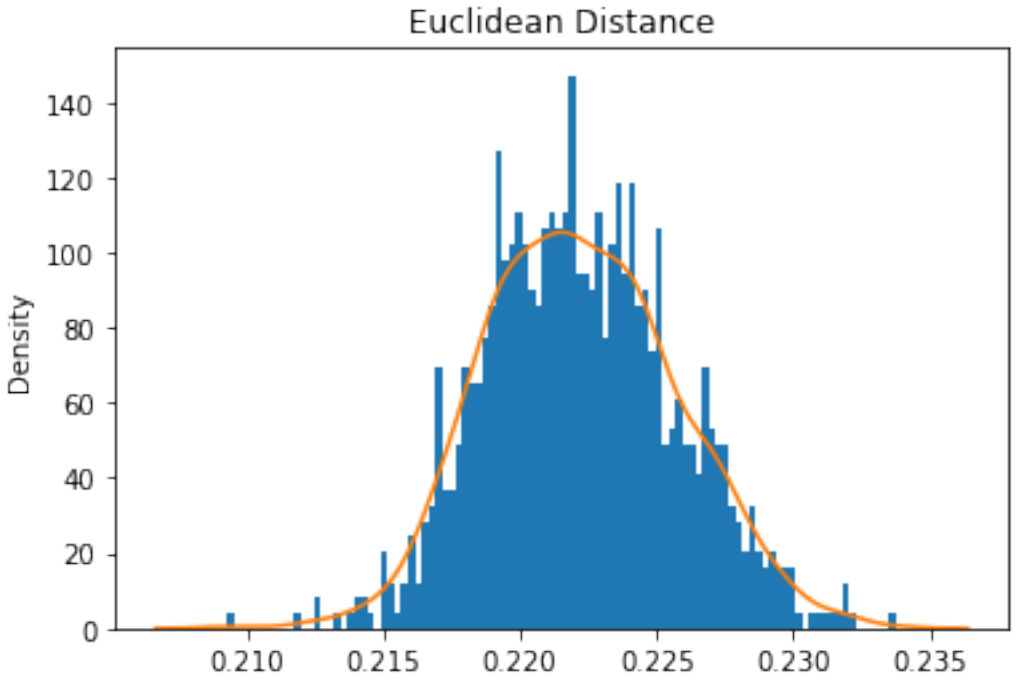
Mean Square Error: 0.0004939509893766551



Mean Absolute Error: 0.016855388164251927

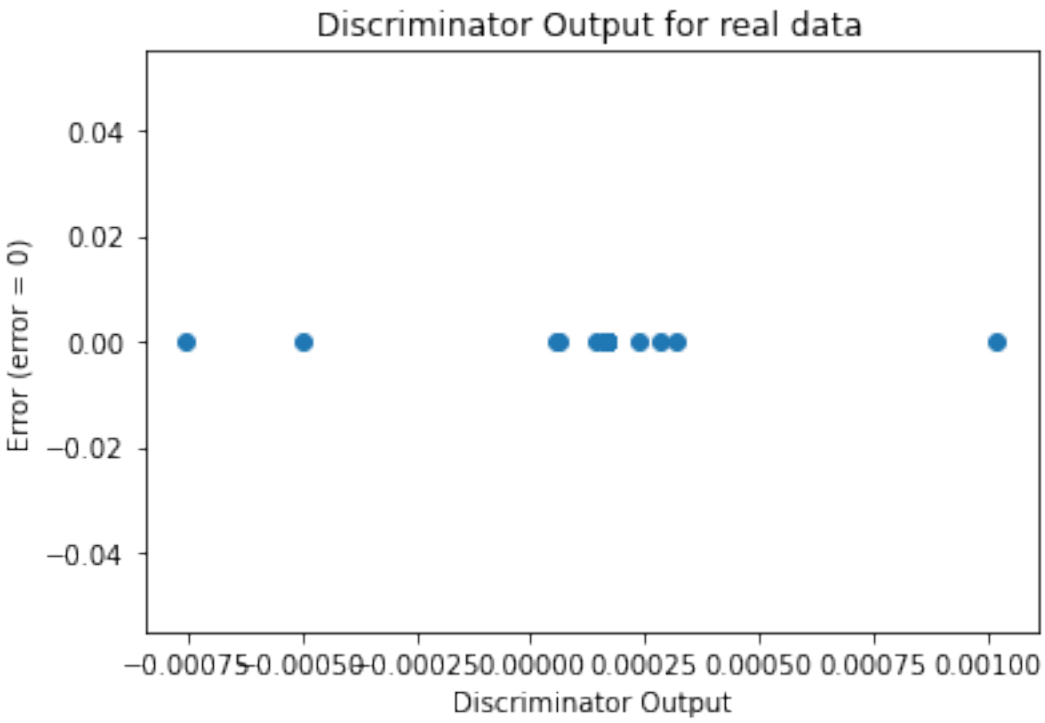


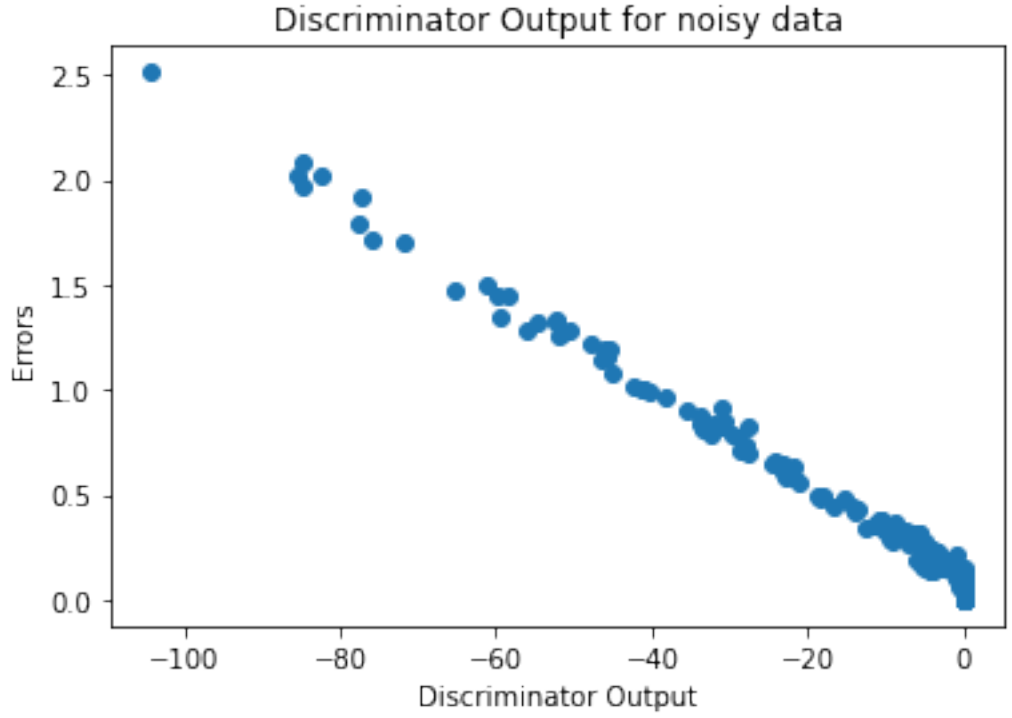
```
Mean Manhattan Distance: 1.6855388164251925
```




```
Mean Euclidean Distance: 0.22222257204195964
```

[13]:
```
sanityChecks.discProbVsError(real_dataset,discriminator,device)
```


Discriminator Output for real data
Error (error = 0)
0.04
0.02
0.00
−0.02
−0.04
−0.00075
−0.00050
−0.00025
0.00000
0.00025
0.00050
0.00075
0.00100
Discriminator Output

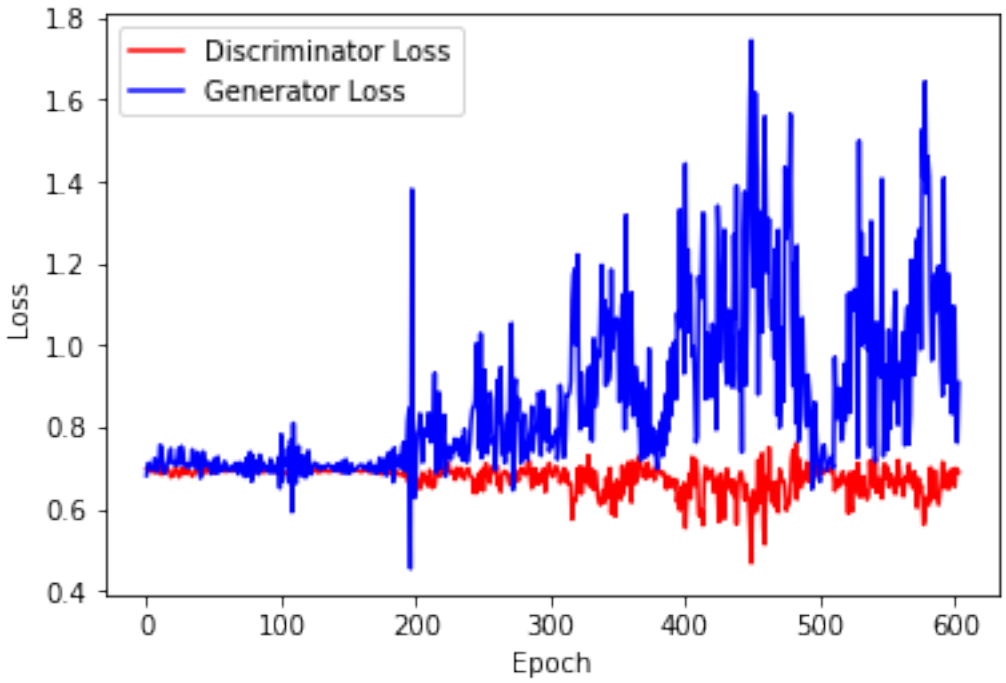

Discriminator Output for noisy data
Errors
2.5
2.0
1.5
1.0
0.5
0.0
−100
−80
−60
−40
−20
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

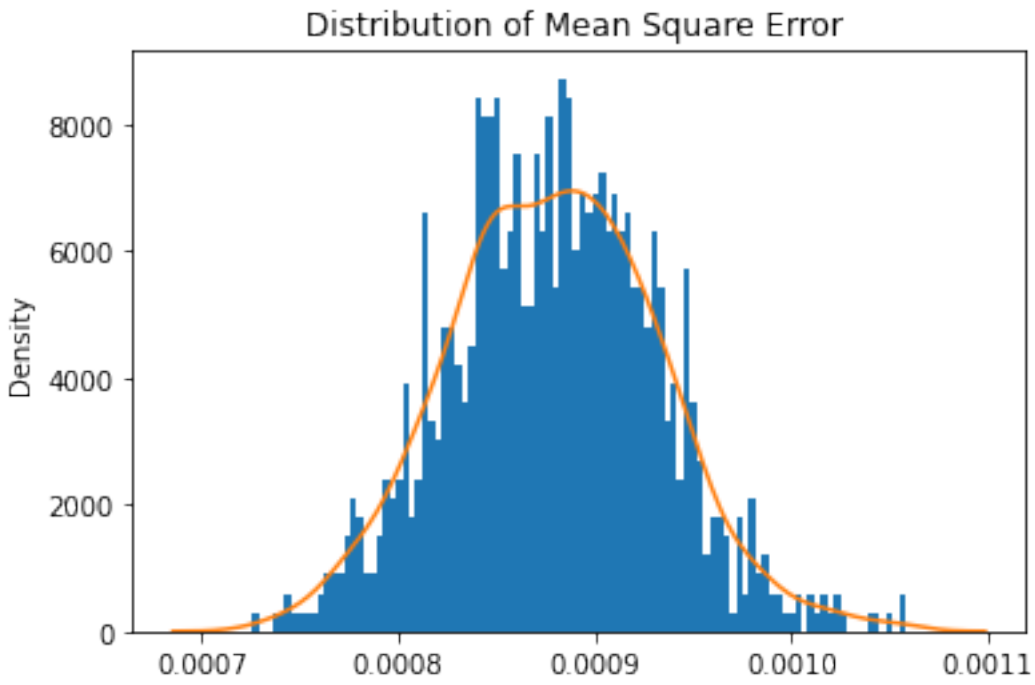
[14]:
```
generator = network.Generator(n_features+2)
discriminator = network.Discriminator(n_features+2)
criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(generator.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(discriminator.parameters(), lr=0.01, betas=(0.5, 0.
↪999))
```

[15]:
```
train_test.
↪training_GAN_2(discriminator,generator,disc_opt,gen_opt,real_dataset,batch_size,error,crite
```

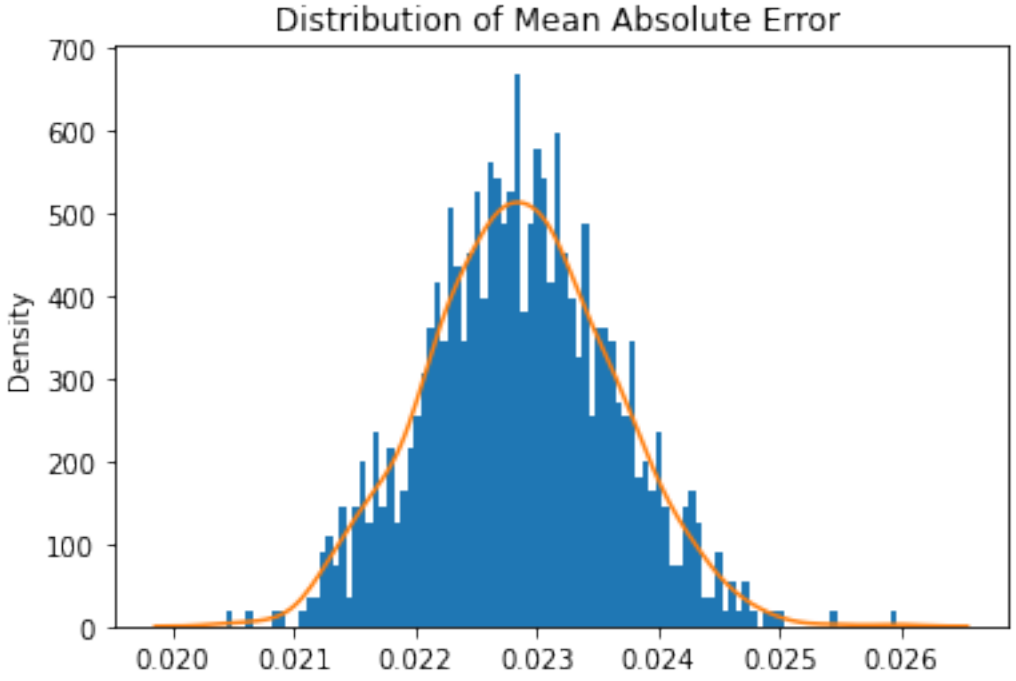
```
Number of epochs needed 302
```



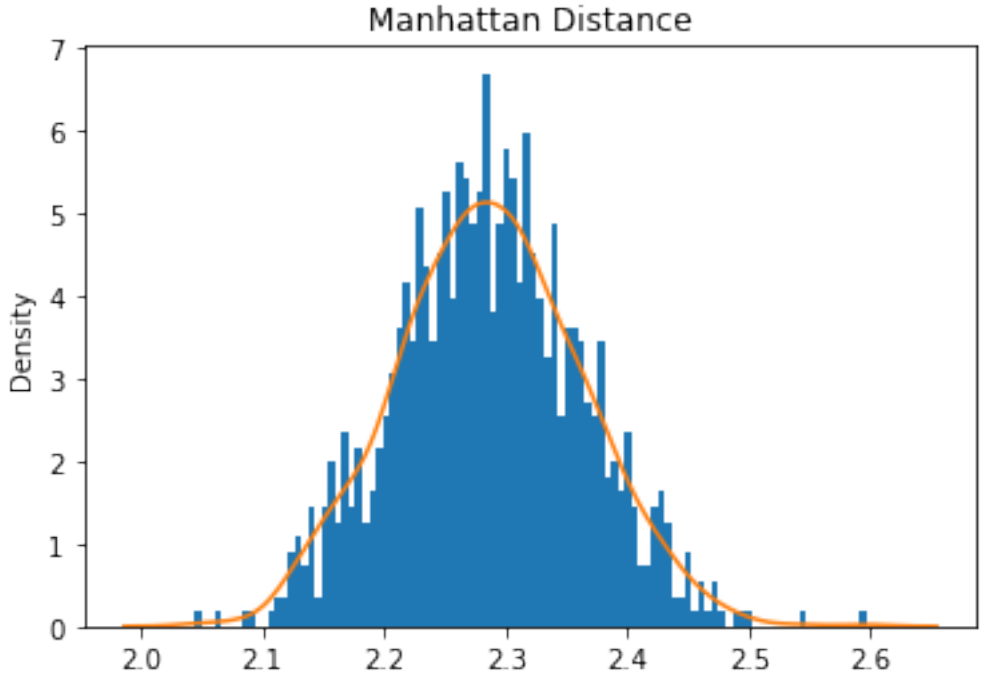
[16]:
```
train_test.test_generator(generator,real_dataset,device)
```



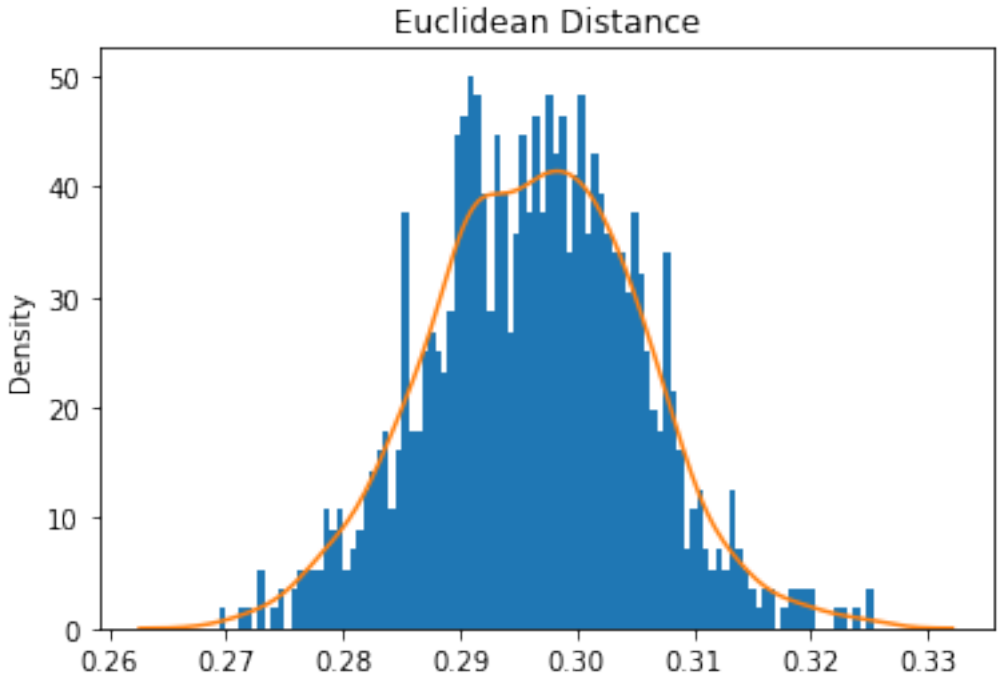

```
Mean Square Error: 0.0008791809438956094
```



```
Mean Absolute Error: 0.022863019734406843
```




```
Mean Manhattan Distance: 2.2863019734406844
```



```
Mean Euclidean Distance: 0.2963730244581927
```

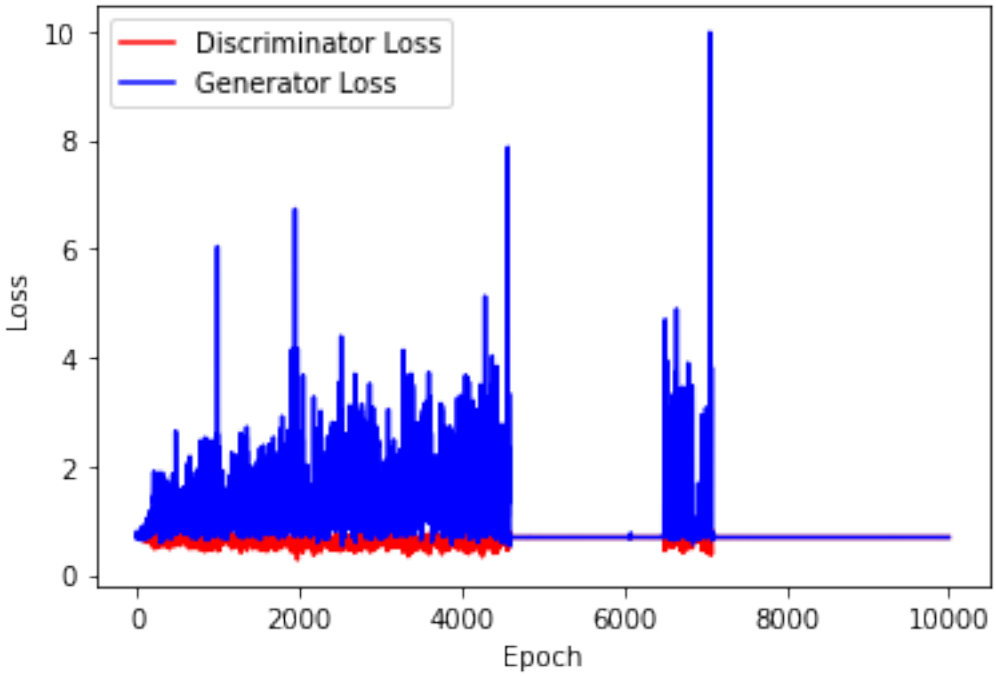
2 ABC GAN Model

2.0.1 Training the network

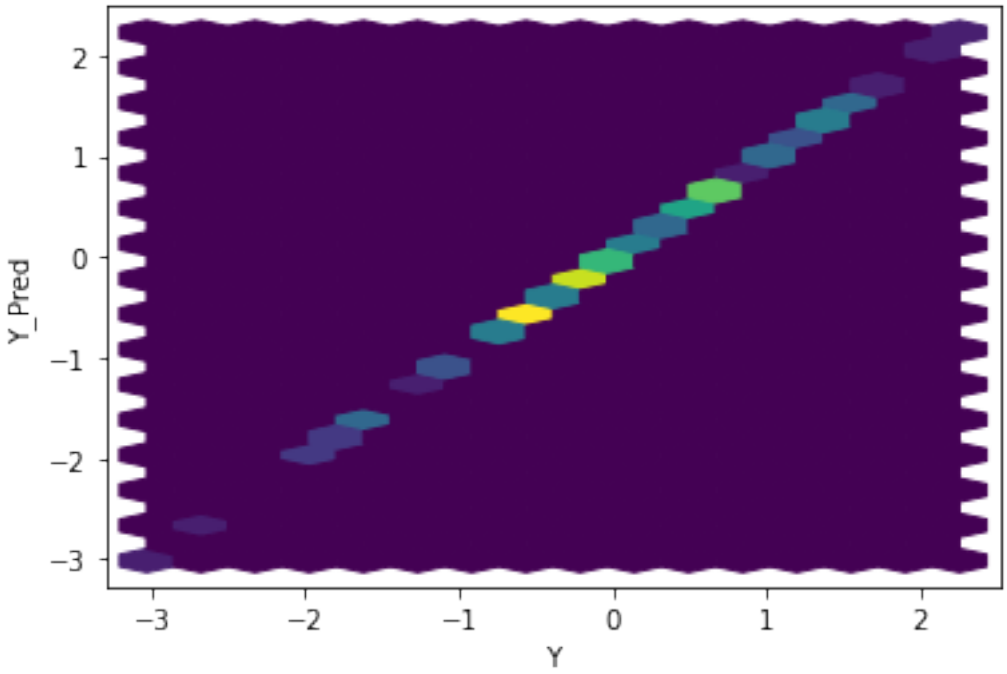
Training ABC-GAN for n_epochs number of epochs

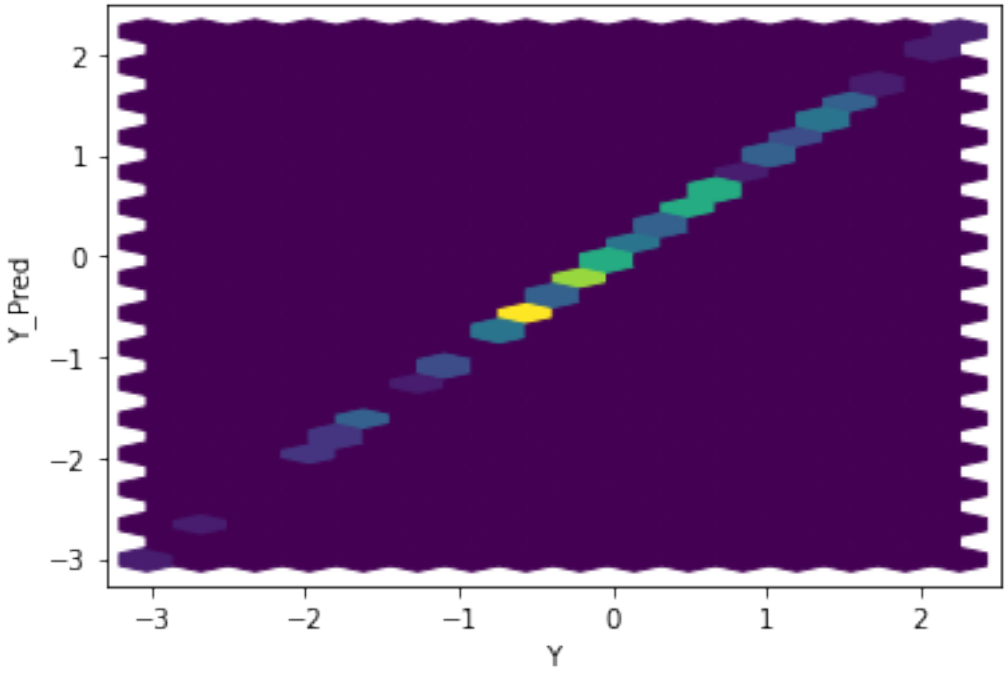
[17]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

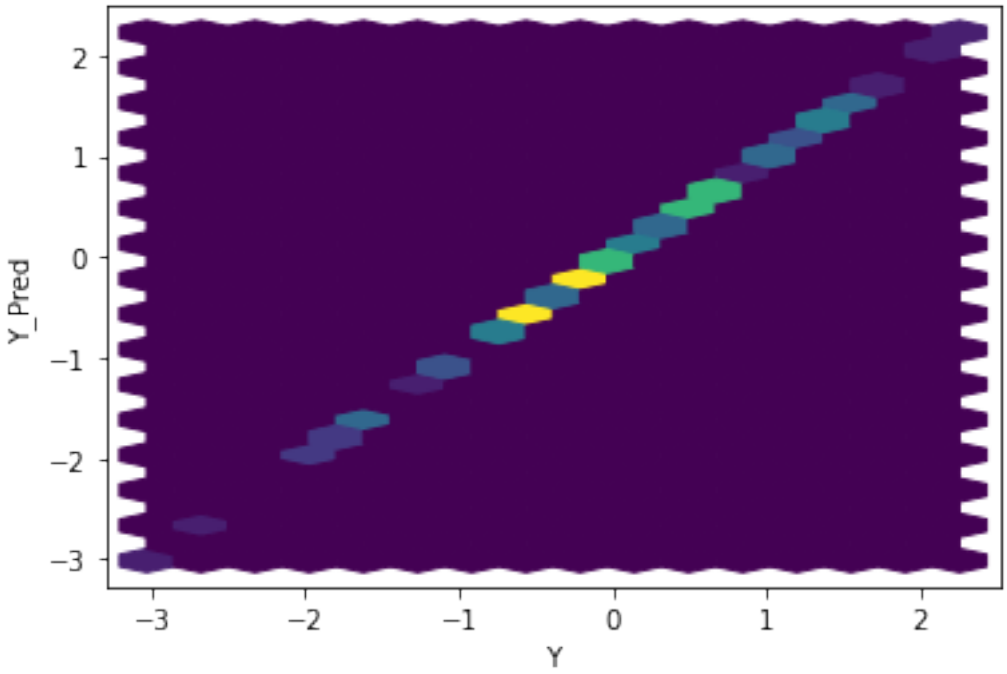
[18]:
```
ABC_train_test.training_GAN(disc, gen,disc_opt,gen_opt,real_dataset,␣
↪batch_size, n_epochs,criterion,coeff,mean,variance,device)
```

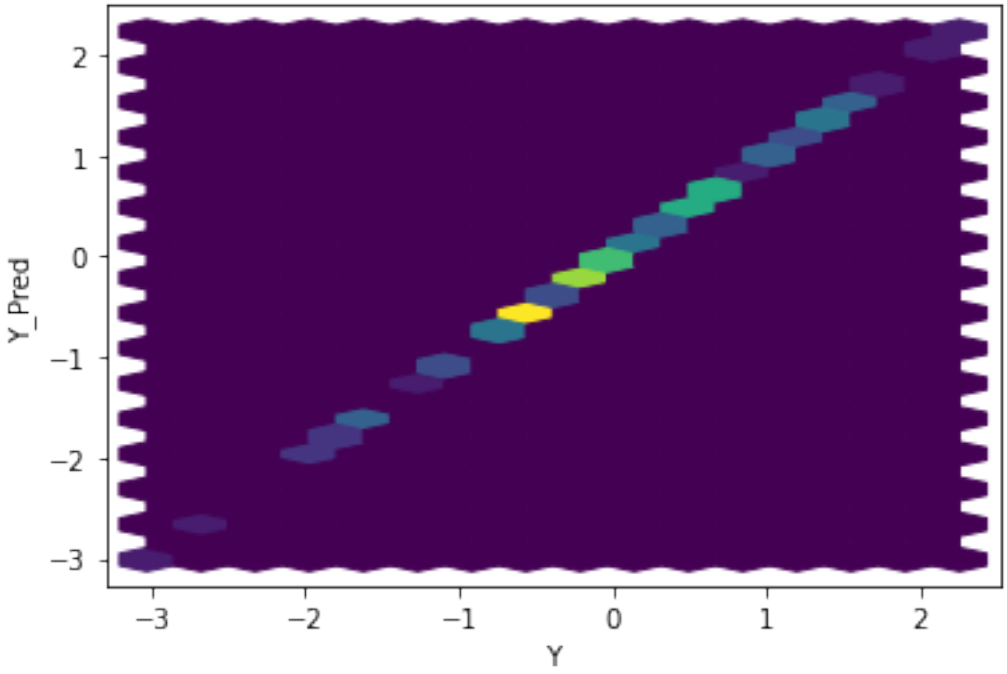


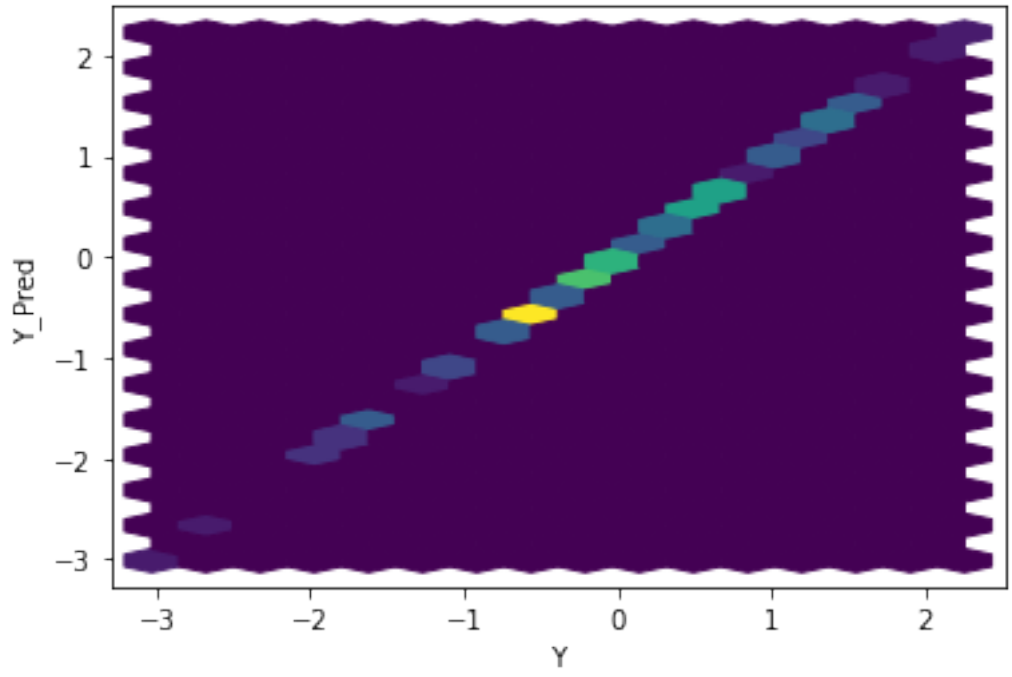
[19]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```

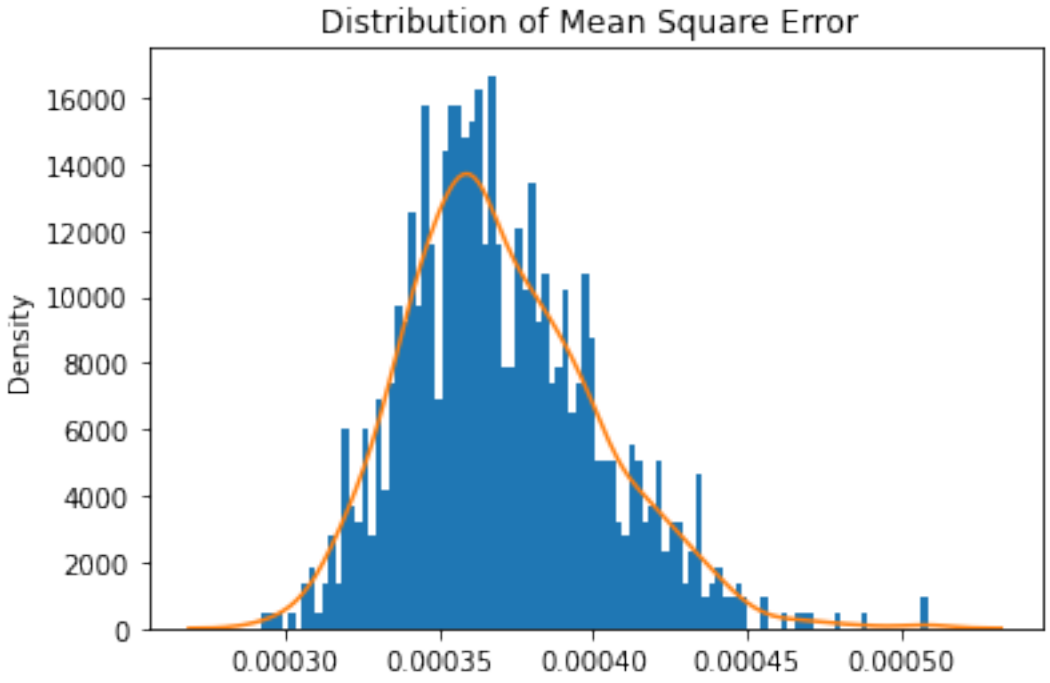

Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

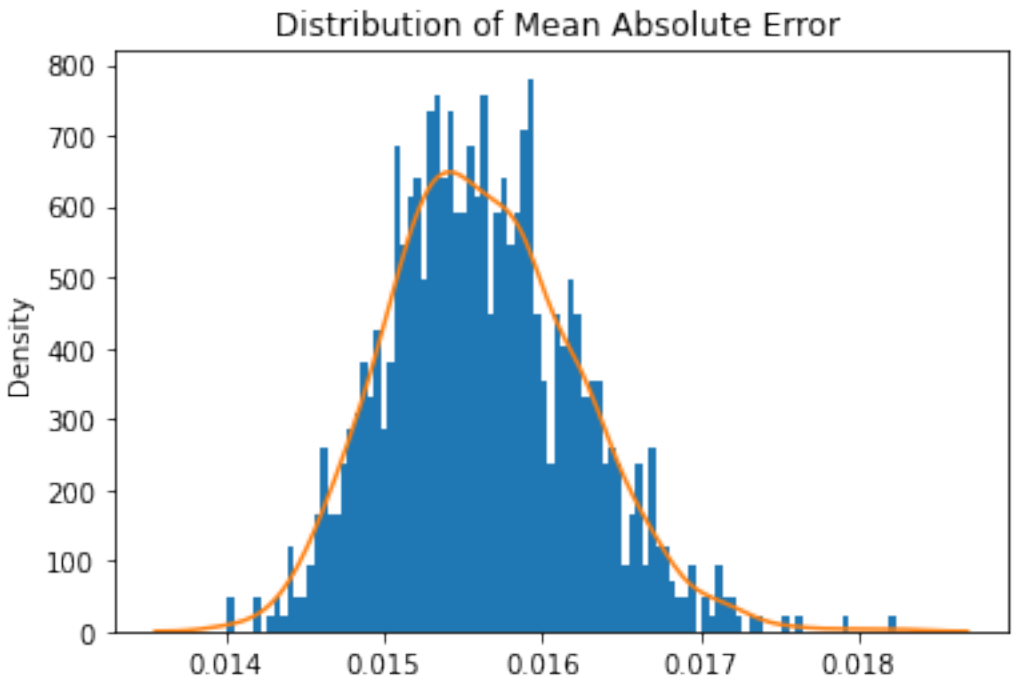

Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

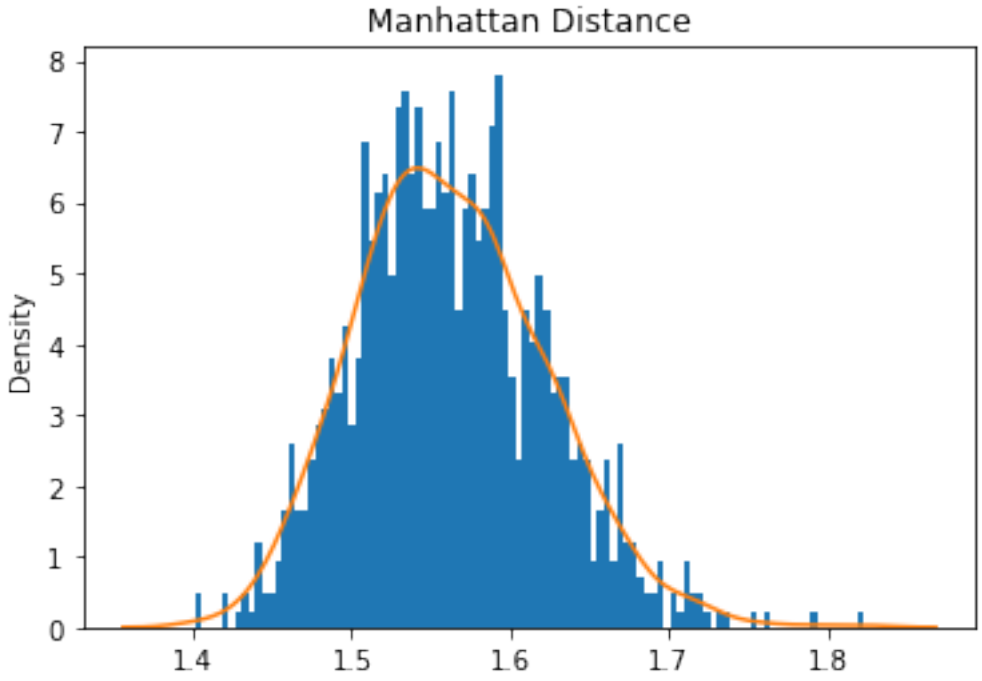




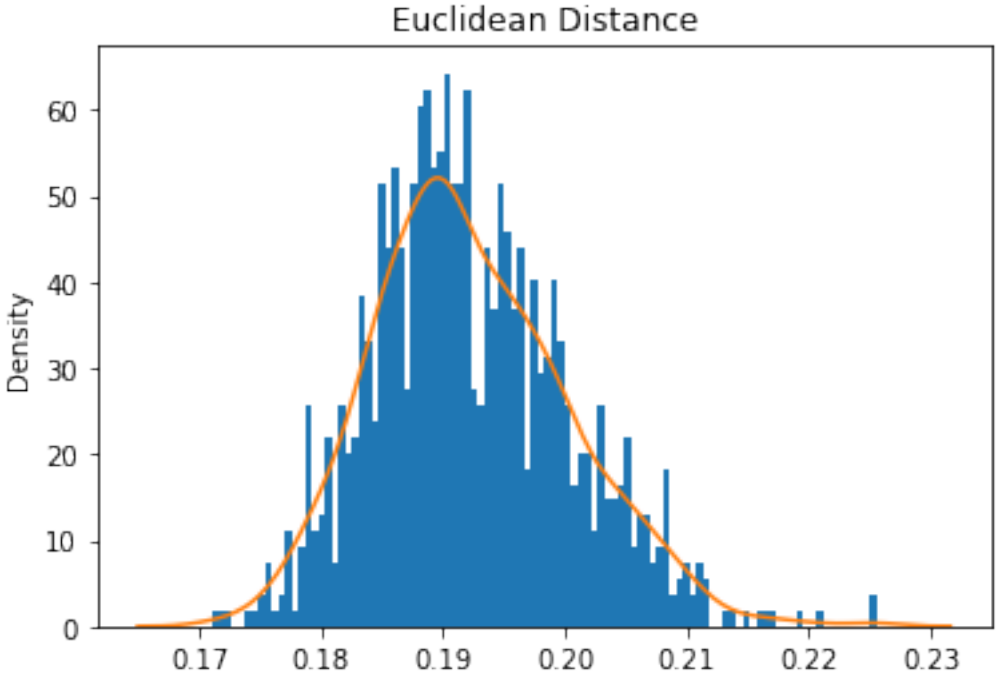

```
Mean Square Error: 0.00037049497909656047
```




```
Mean Absolute Error: 0.01562638223710237
Mean Manhattan Distance: 1.562638223710237
```

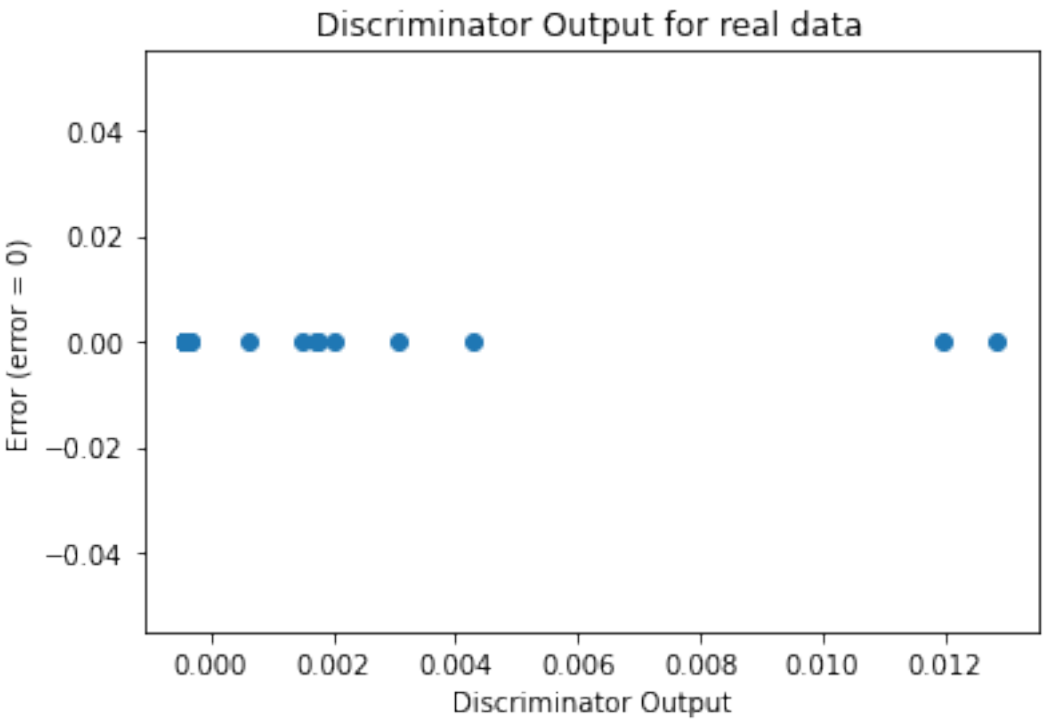


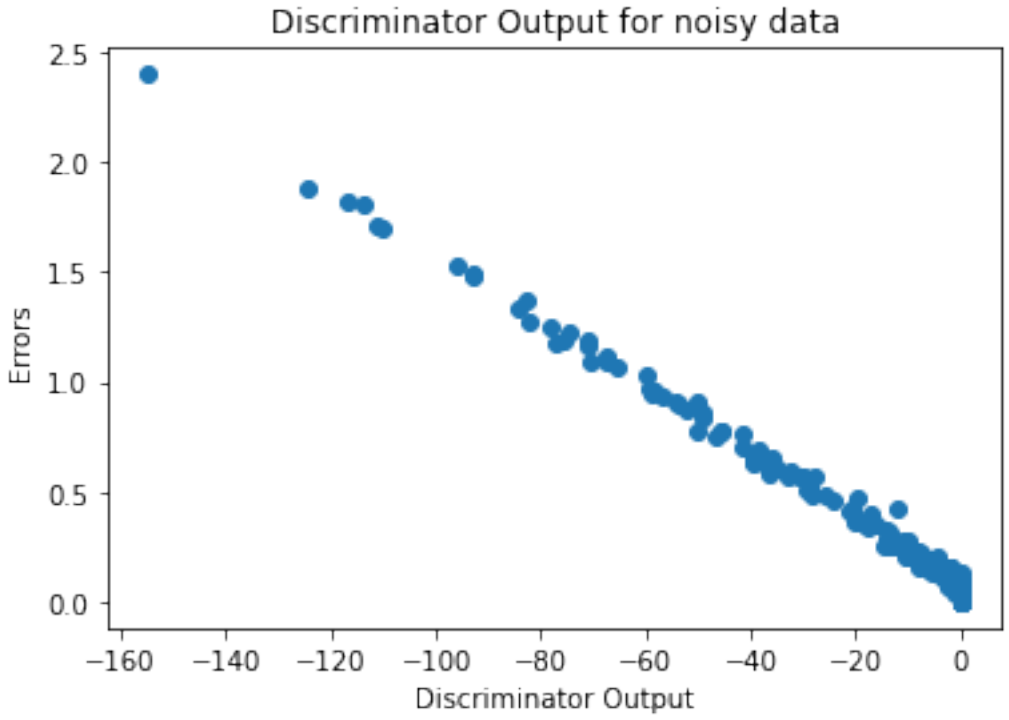
```
Mean Euclidean Distance: 0.1923120538549535
```



Sanity Checks

[20]:
```
sanityChecks.discProbVsError(real_dataset,disc,device)
```


Discriminator Output for real data
Error (error = 0)
0.04
0.02
0.00
−0.02
−0.04
0.000
0.002
0.004
0.006
0.008
0.010
0.012
Discriminator Output

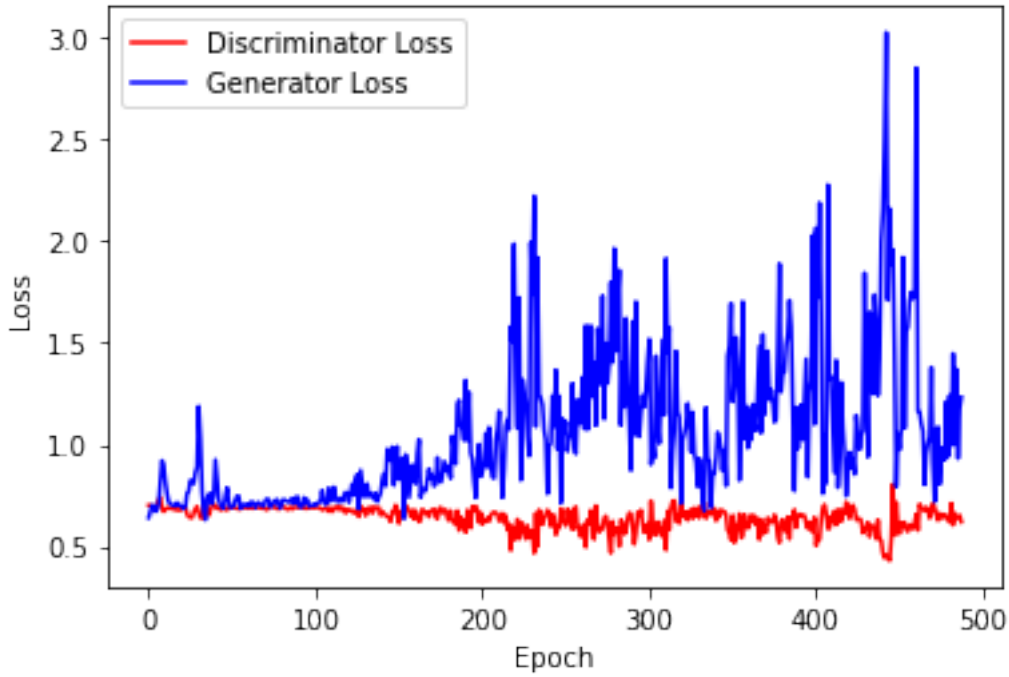

Discriminator Output for noisy data
Errors
2.5
2.0
1.5
1.0
0.5
0.0
−160
−140
−120
−100
−80
−60
−40
−20
0
Discriminator Output

Training GAN until mse of y_pred is > 0.1 or n_epochs < 30000

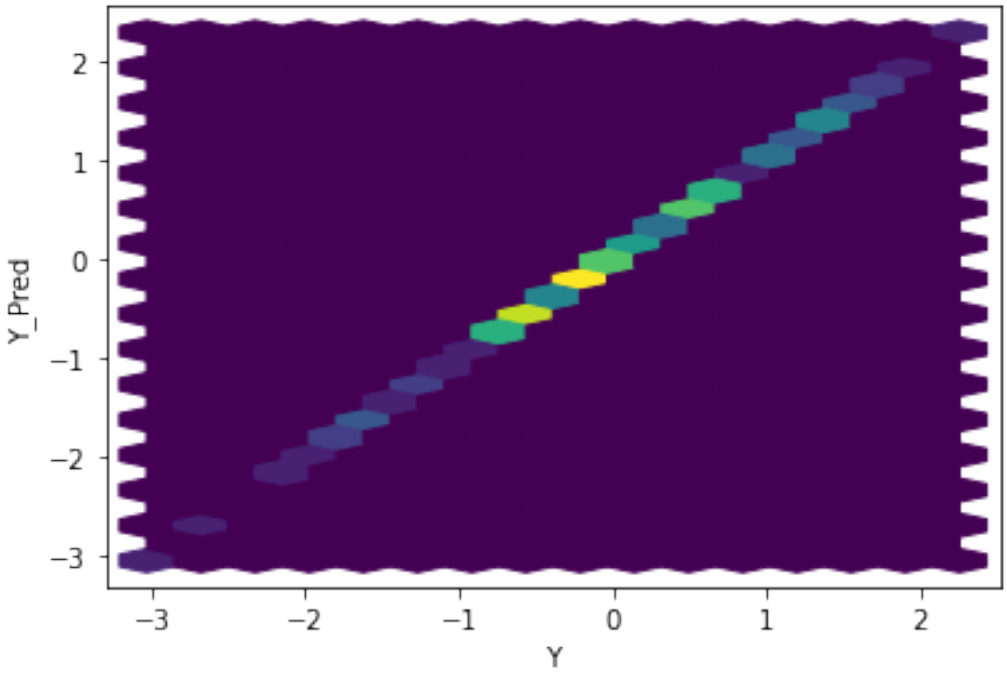
[21]:
```
gen = network.Generator(n_features+2)
disc = network.Discriminator(n_features+2)

criterion = torch.nn.BCEWithLogitsLoss()
gen_opt = torch.optim.Adam(gen.parameters(), lr=0.01, betas=(0.5, 0.999))
disc_opt = torch.optim.Adam(disc.parameters(), lr=0.01, betas=(0.5, 0.999))
```

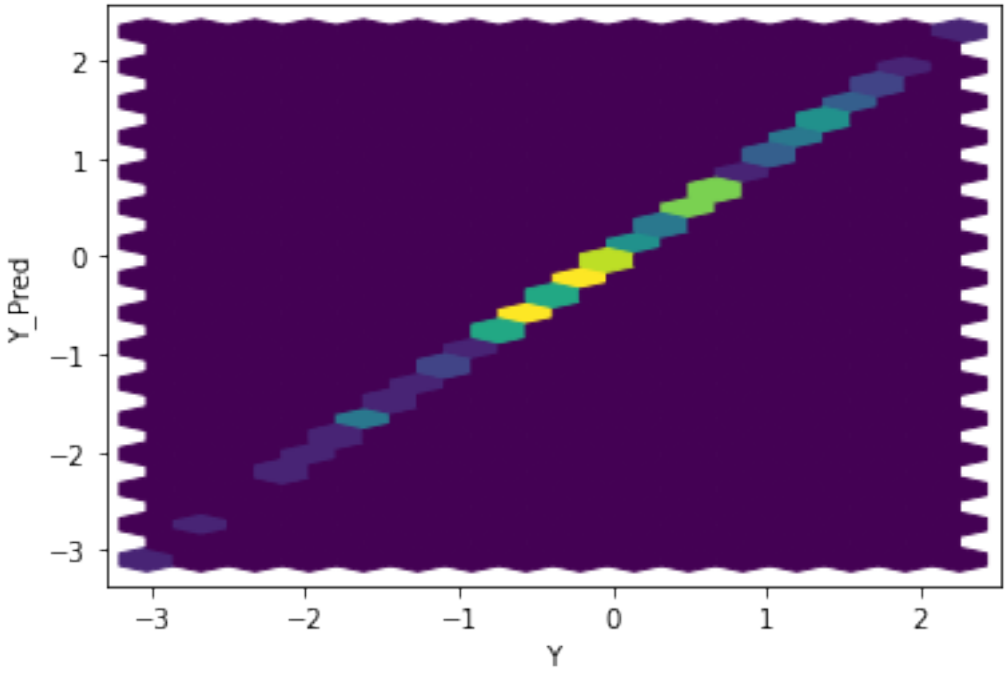
[22]:
```
ABC_train_test.
↪training_GAN_2(disc,gen,disc_opt,gen_opt,real_dataset,batch_size,␣
↪error,criterion,coeff,mean,variance,device)
```

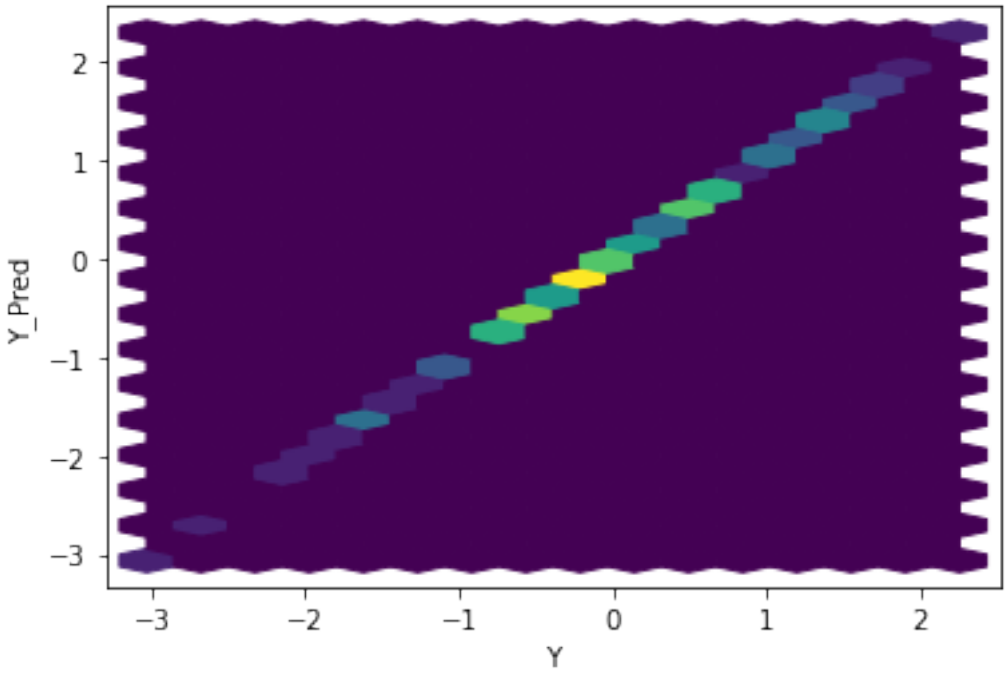
```
Number of epochs 244
```

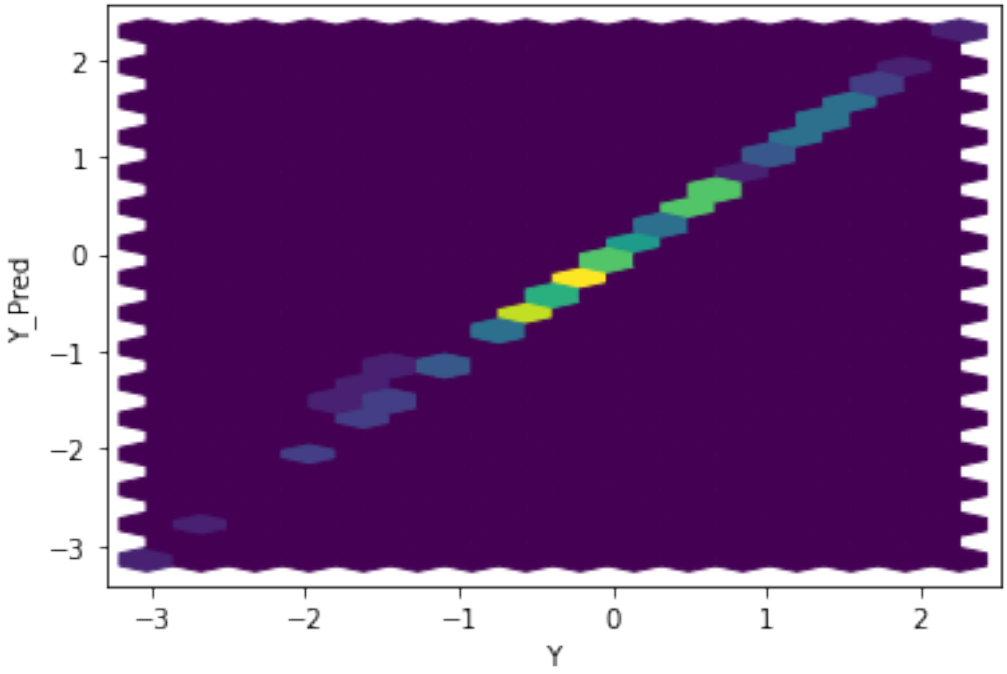


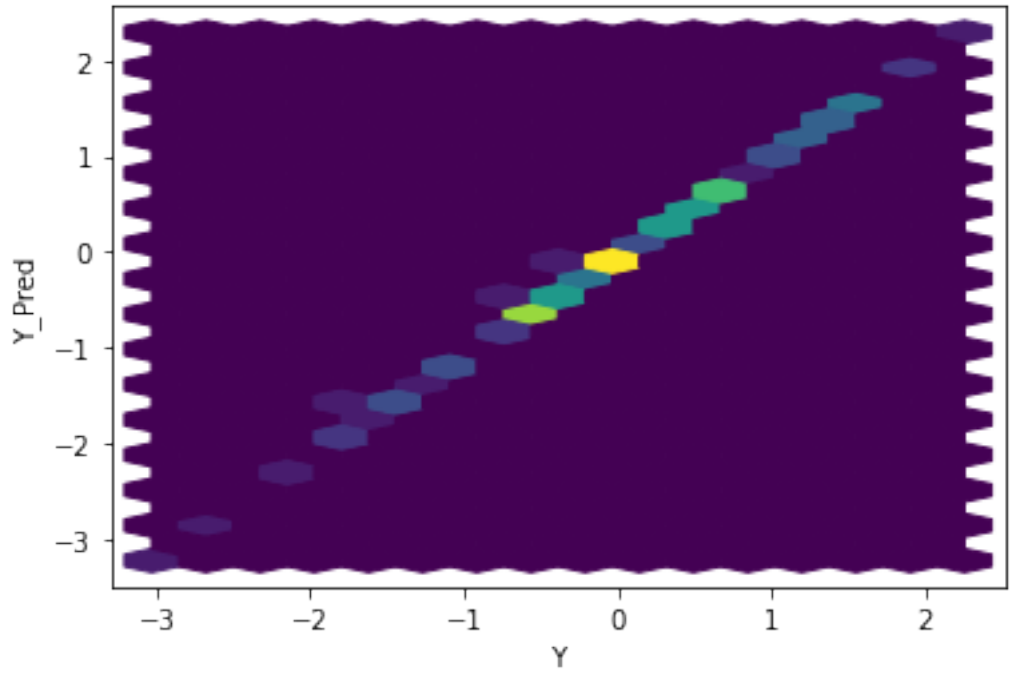
[23]:
```
ABC_train_test.test_generator(gen,real_dataset,coeff,mean,variance,device)
```

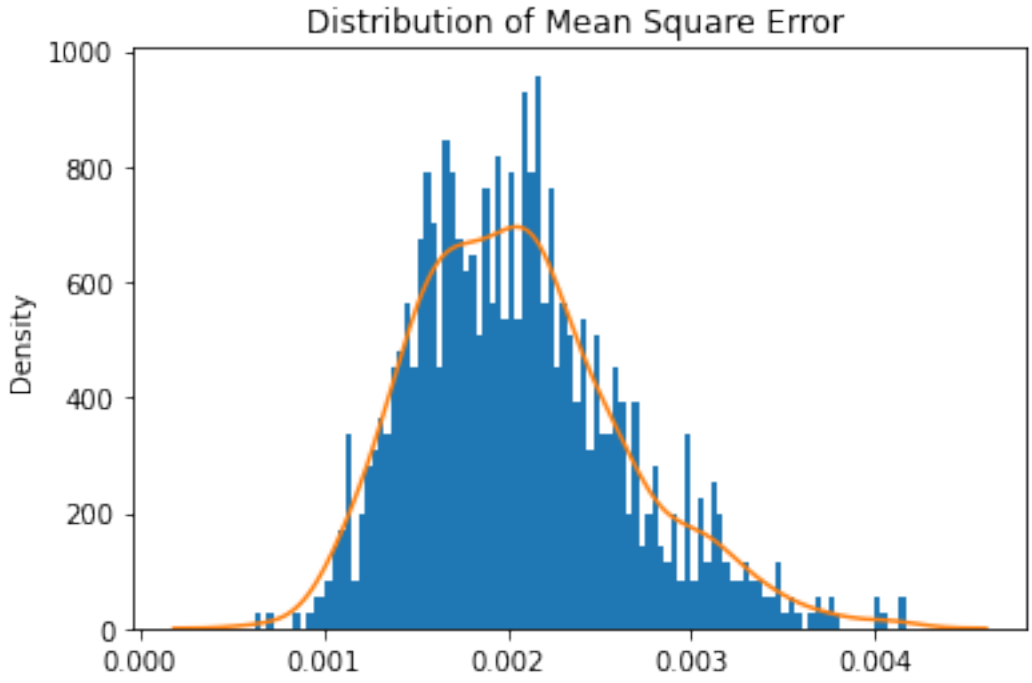

Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

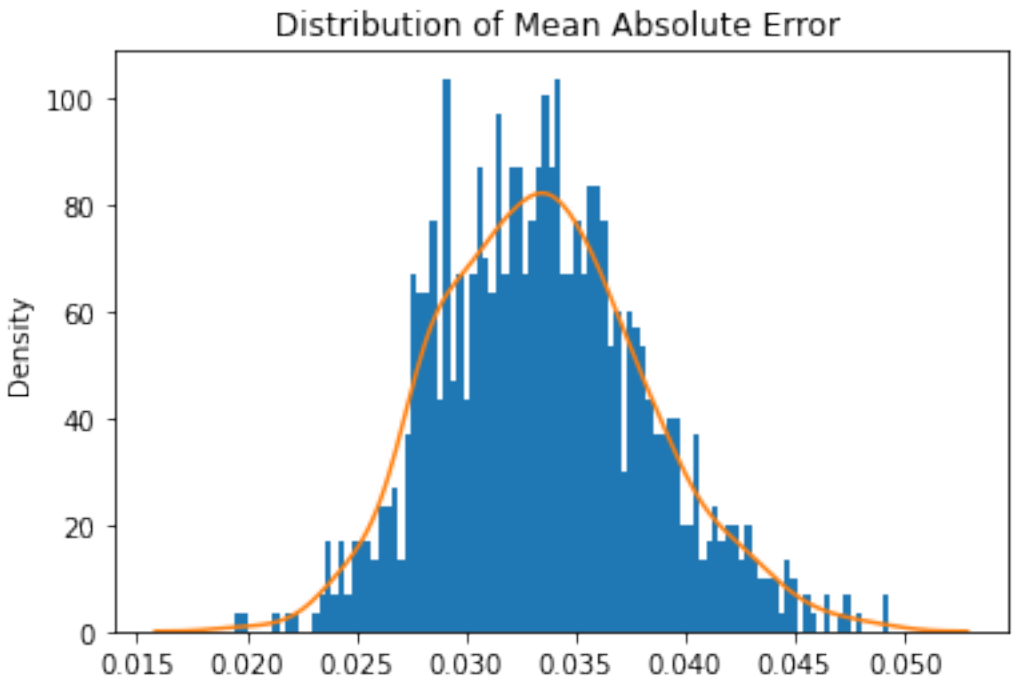

Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2


Y_Pred
Y
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

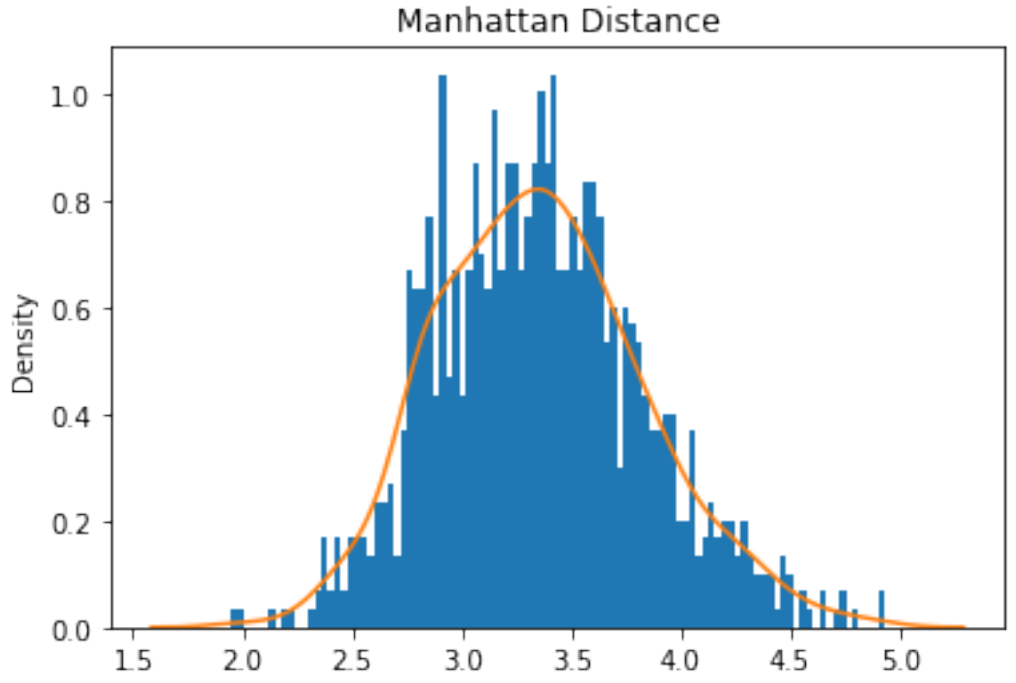




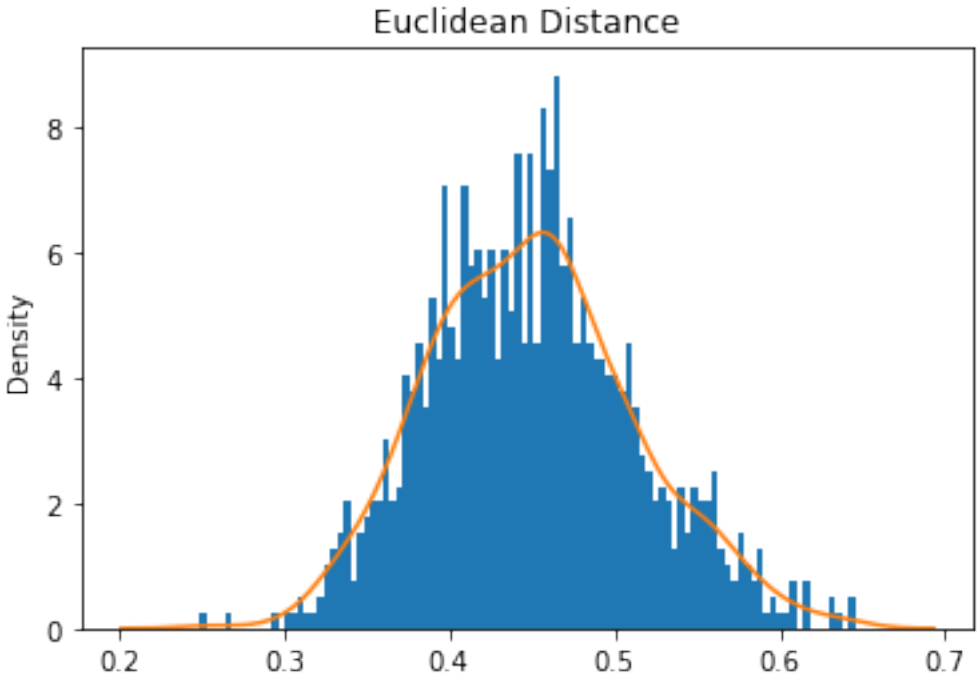

```
Mean Square Error: 0.0020636950090907303
```




```
Mean Absolute Error: 0.033541463671839446
Mean Manhattan Distance: 3.3541463671839447
```



```
Mean Euclidean Distance: 0.4499181558883759
```



[ ]: