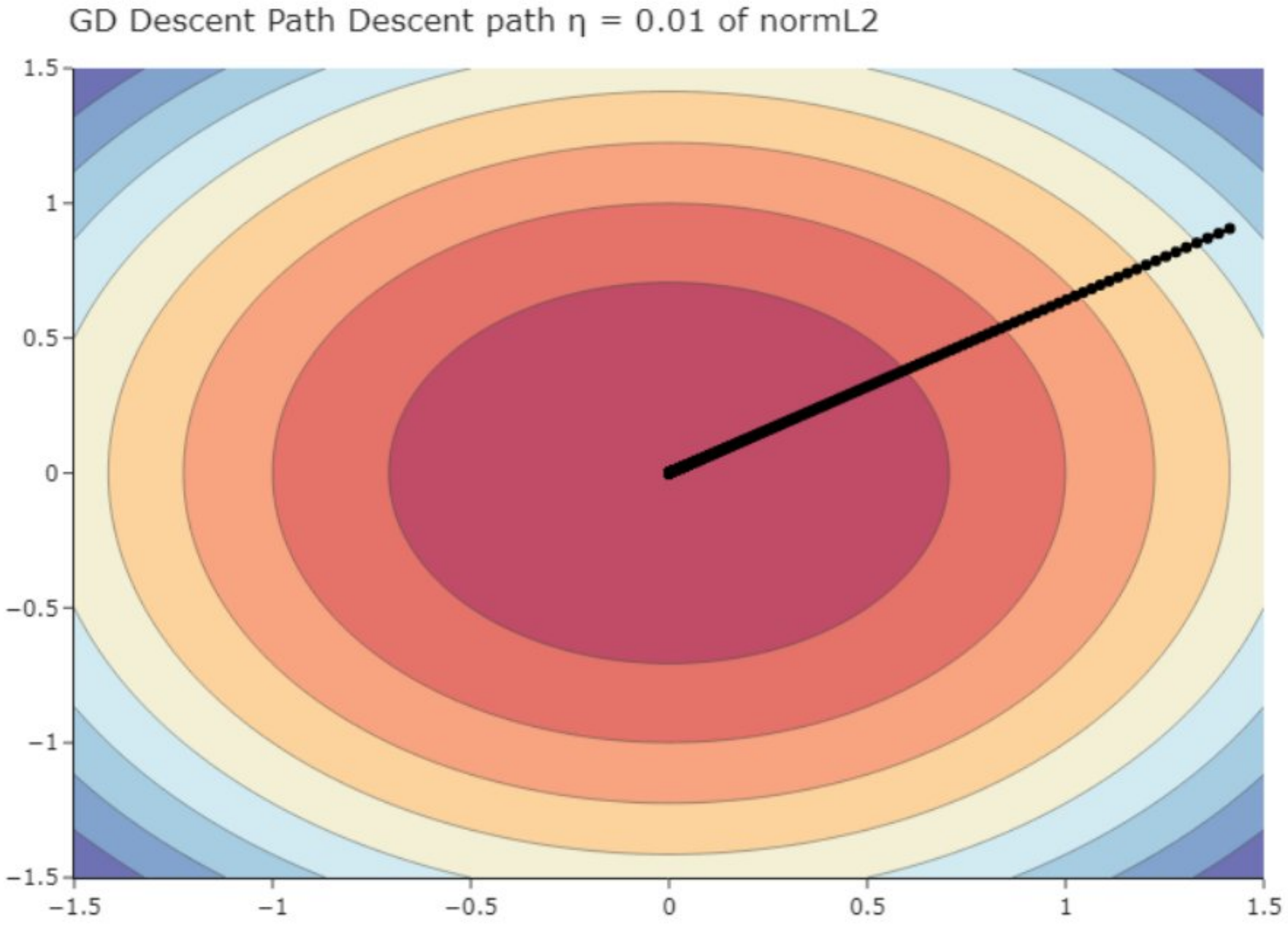
Practical part:

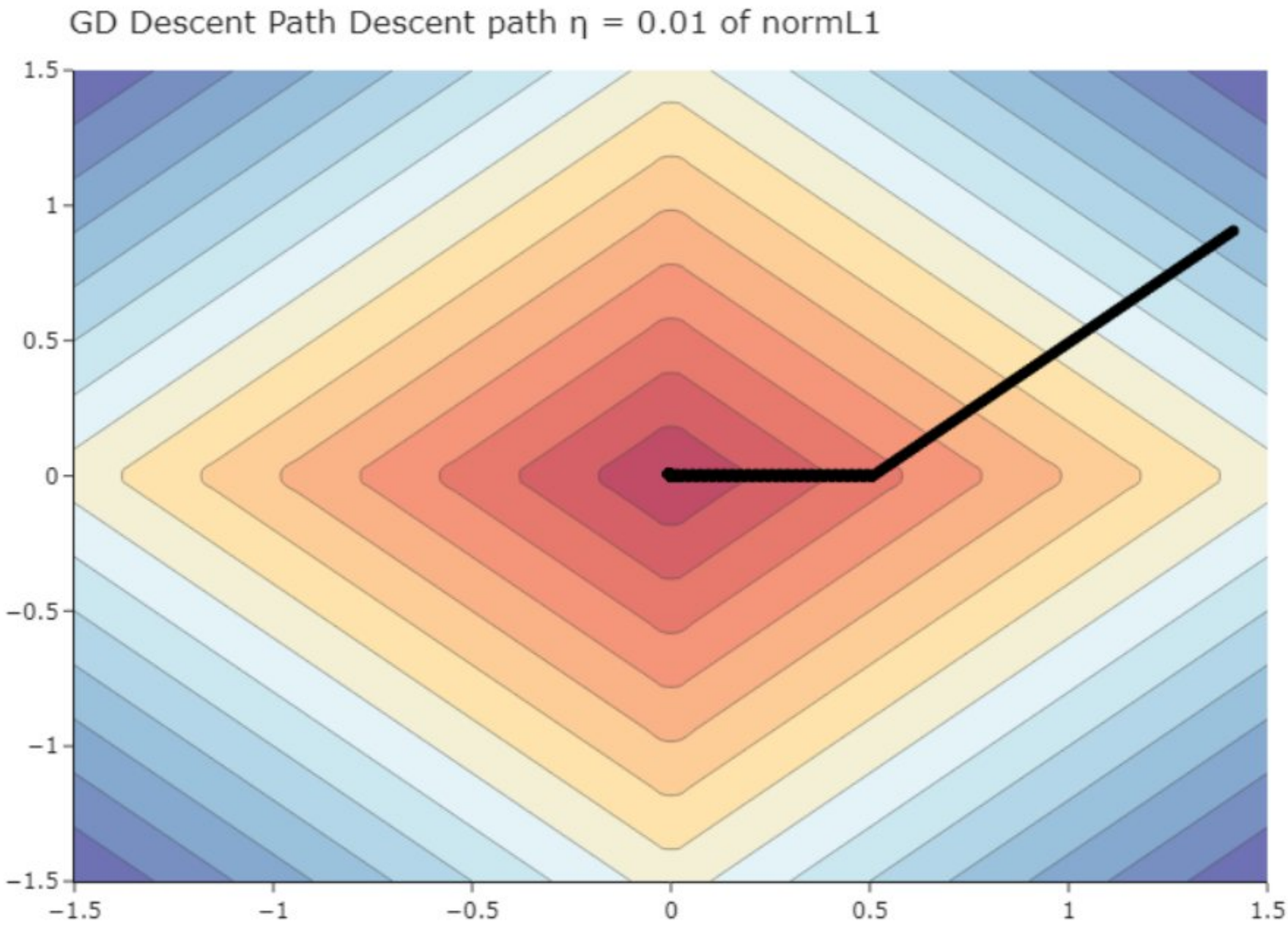
(The answers for theoretical part attached below)

Part 2.1.1 Comparing Fixed learning rates:

Q1: The descent path for $\eta = 0.01$ and explain the differences seen between the L1 and L2 modules:



Scanned with CamScanner

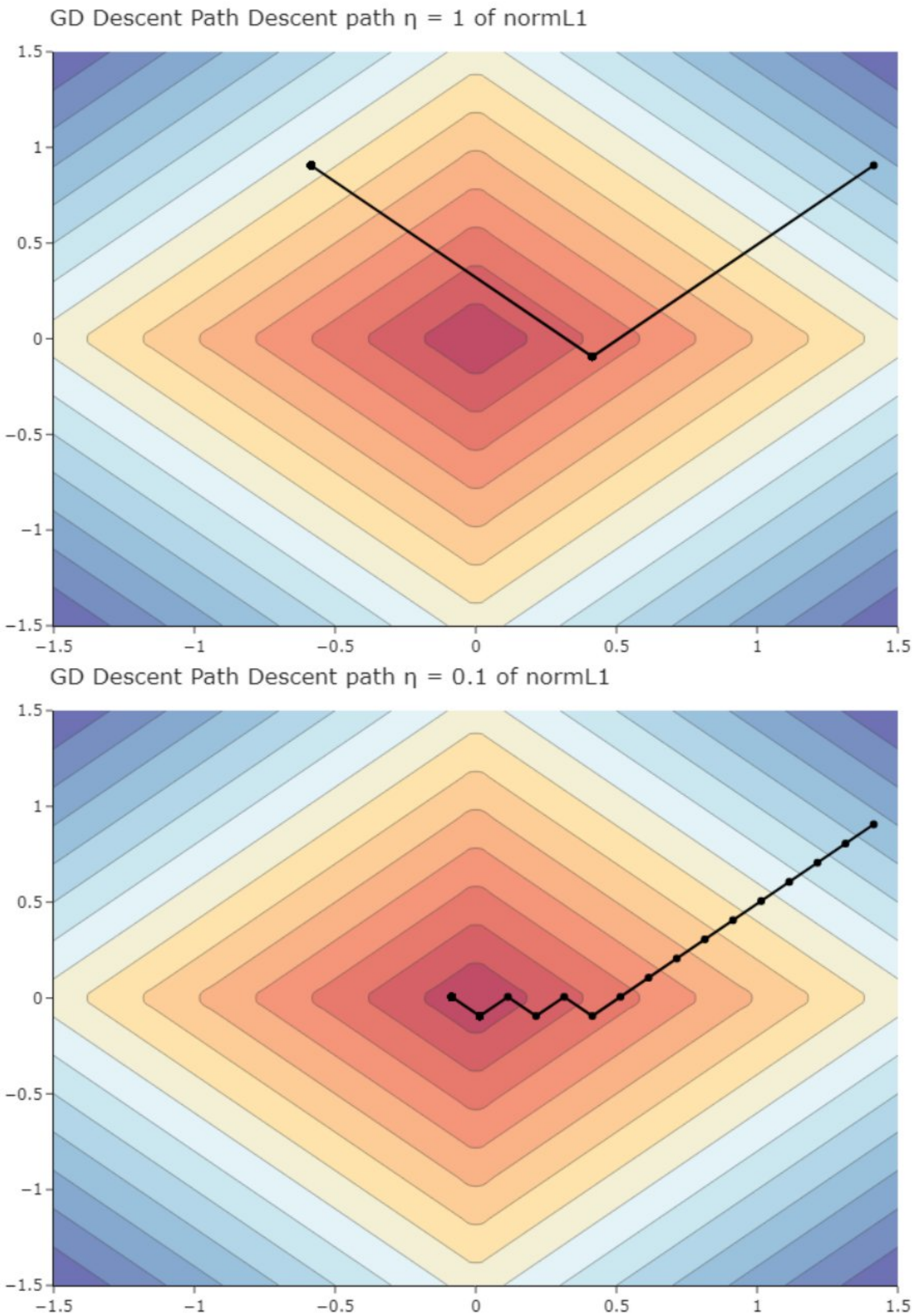


Exploitation:

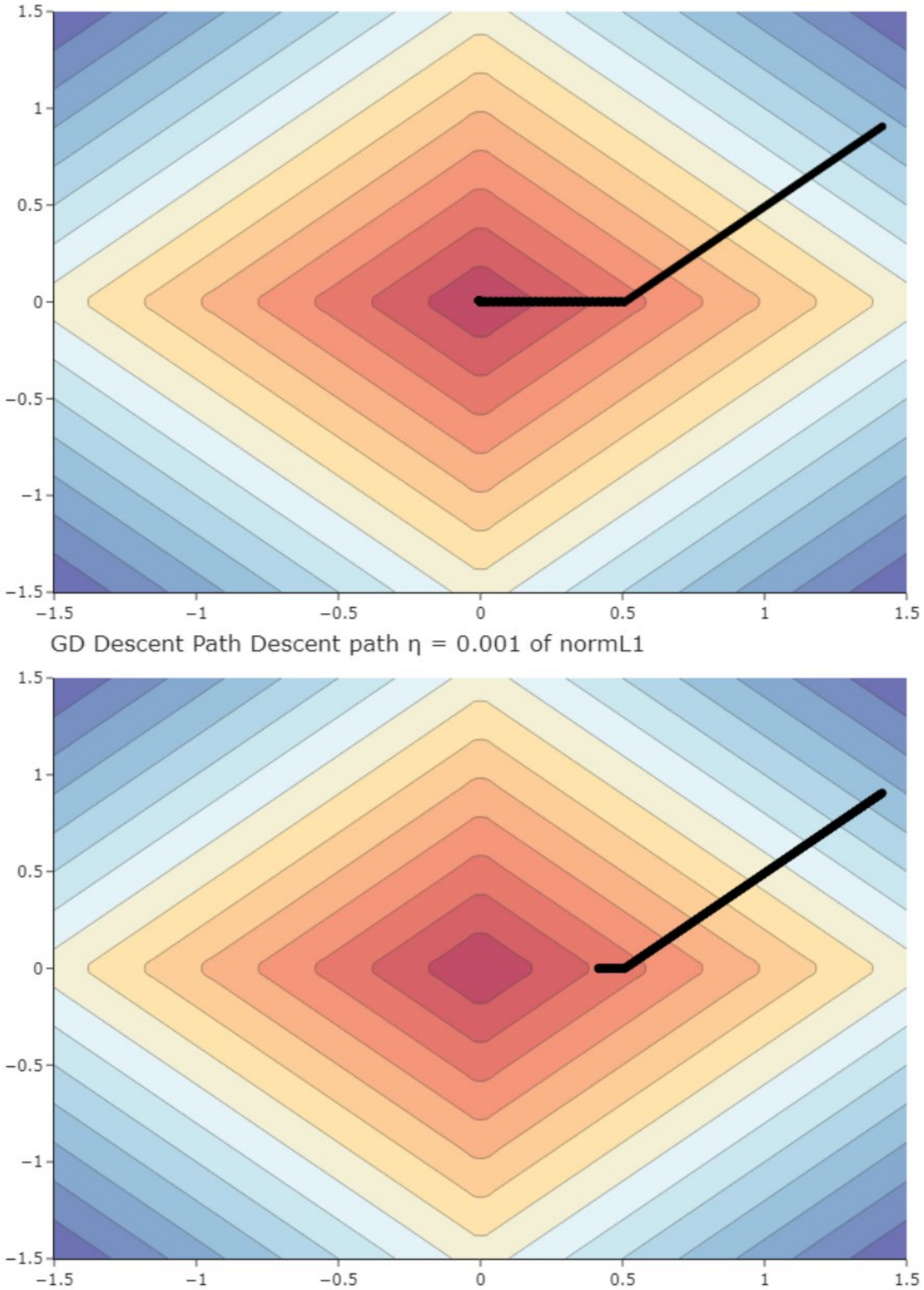
The approaching learning rate to the minimal norm is different! In L2 the pace is relatively constant which is expressed in a straight line and a smooth descent. On the other hand, in L2 the decline was not smooth and at a certain point the direction of the decline was broken. The difference between them comes from the difference in the gradient functions. While L2 is a smooth and differentiable function its gradient is calculated directly, L1 is neither differentiable nor smooth and a sub-gradient is needed. L1 does not converge to zero compared to L2 which converges to zero.

Scanned with CamScanner

Q2: Description of two phenomena that can be seen in the descent path of the ℓ_1 objective when using GD and a fixed learning rate:



Scanned with CamScanner



Exploitation:

When we used η of size 1 we got too big a jump between iterations, meaning the progression in the size of the gradient is too big! When we used a η of 0.001 the progress was too slow and the number of iterations was not enough to minimize the norm. It can be observed that in different values of η, L1 does not converge

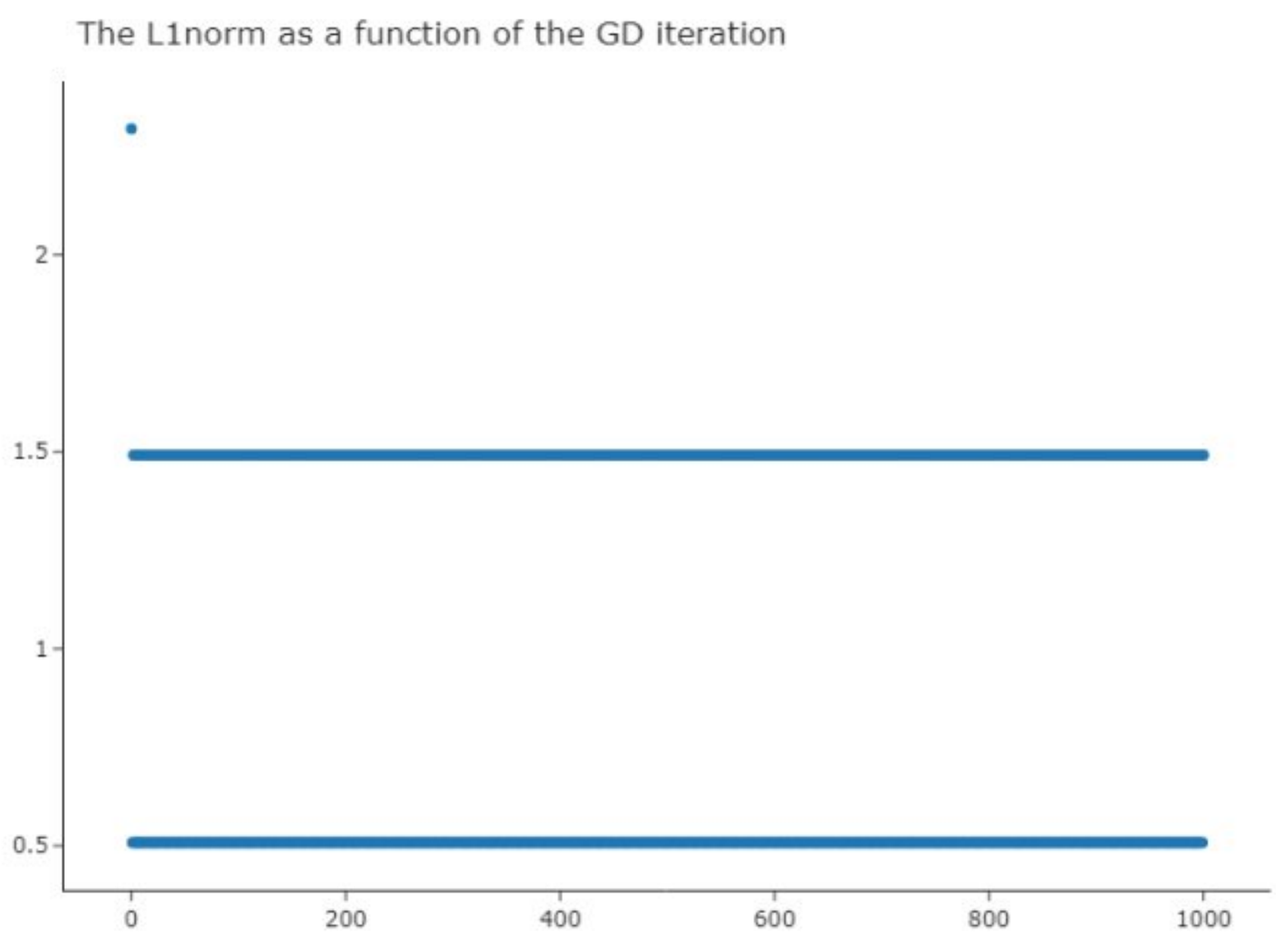
Scanned with CamScanner

to absolute 0. Since it is an L1 norm, we will notice that the ranges of possible norms are in the form of a rhombus.

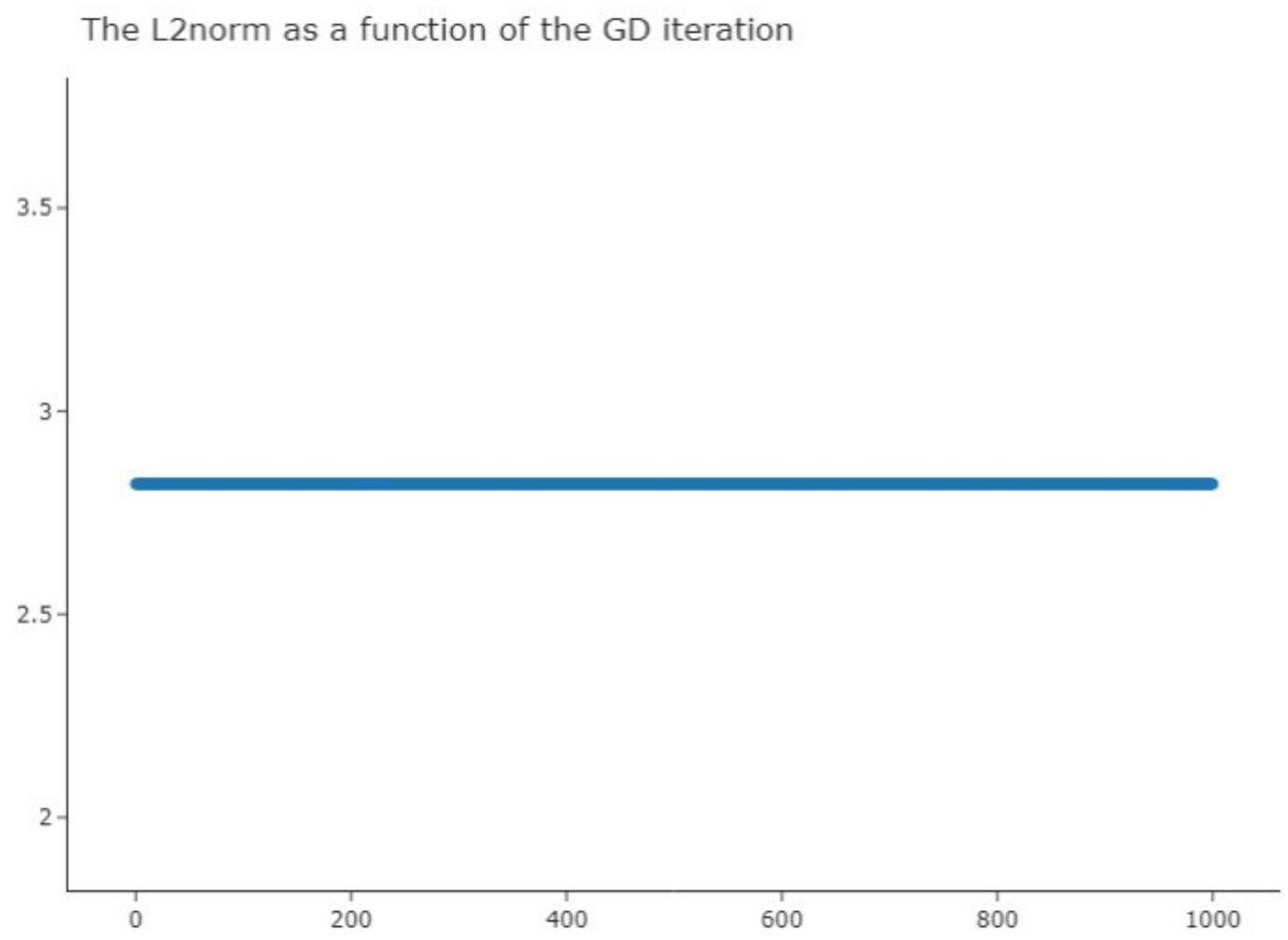
Q3: For each of the modules, the convergence rate for all specified learning rates plots:

(*) For $\eta = 1$

L1:



L2:



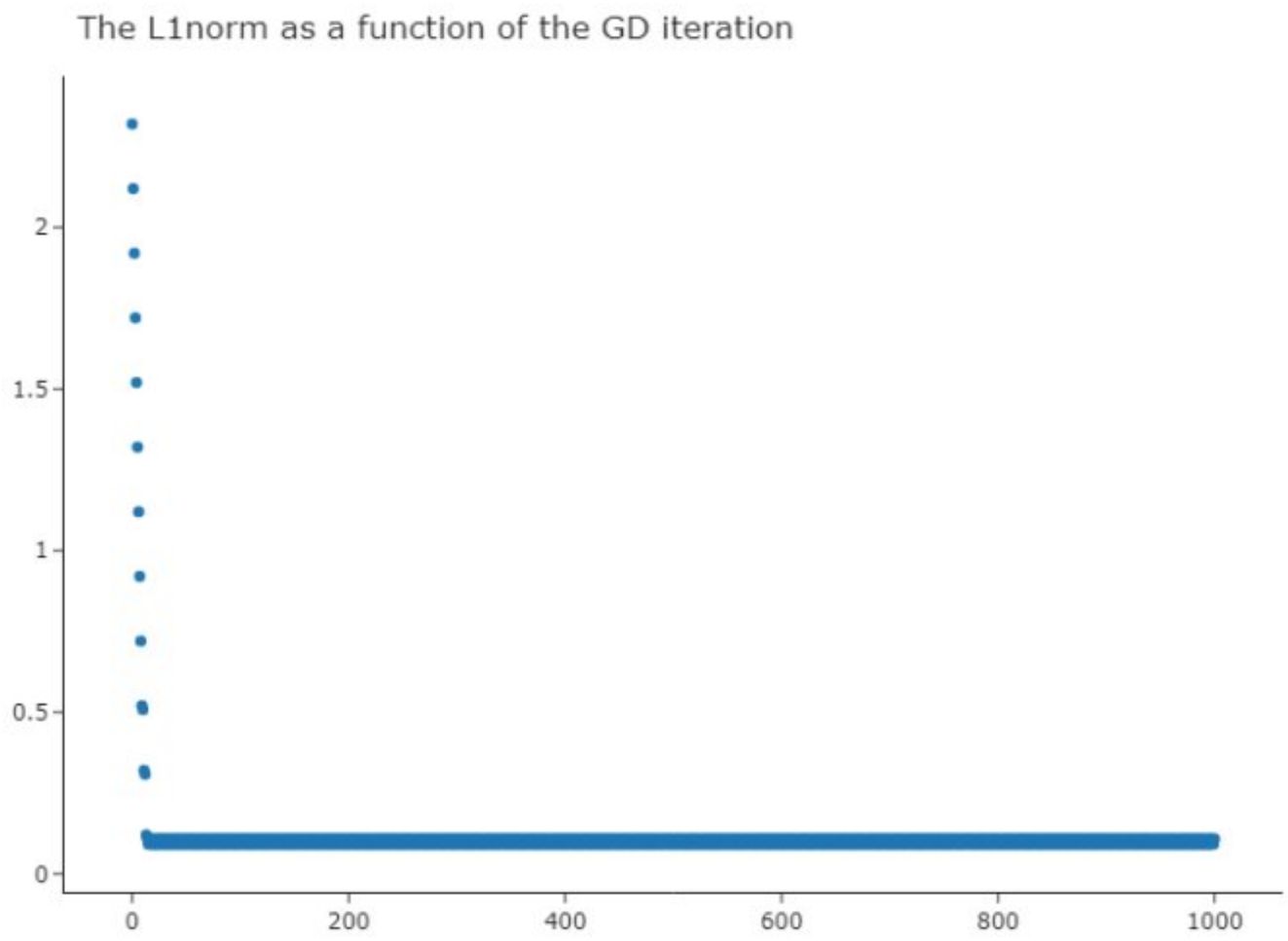
Exploitation:

In L1, the learning constant is high, the jumps between iterations are between 2 fixed points - one with a norm that curves to zero and one with a big norm. From the low point we don't get down because you bend too much in the direction of the gradient. In the L2 norm, the jump between iterations is between two fixed points with the same norm, that is, the norm does not decrease in size during the iterations.

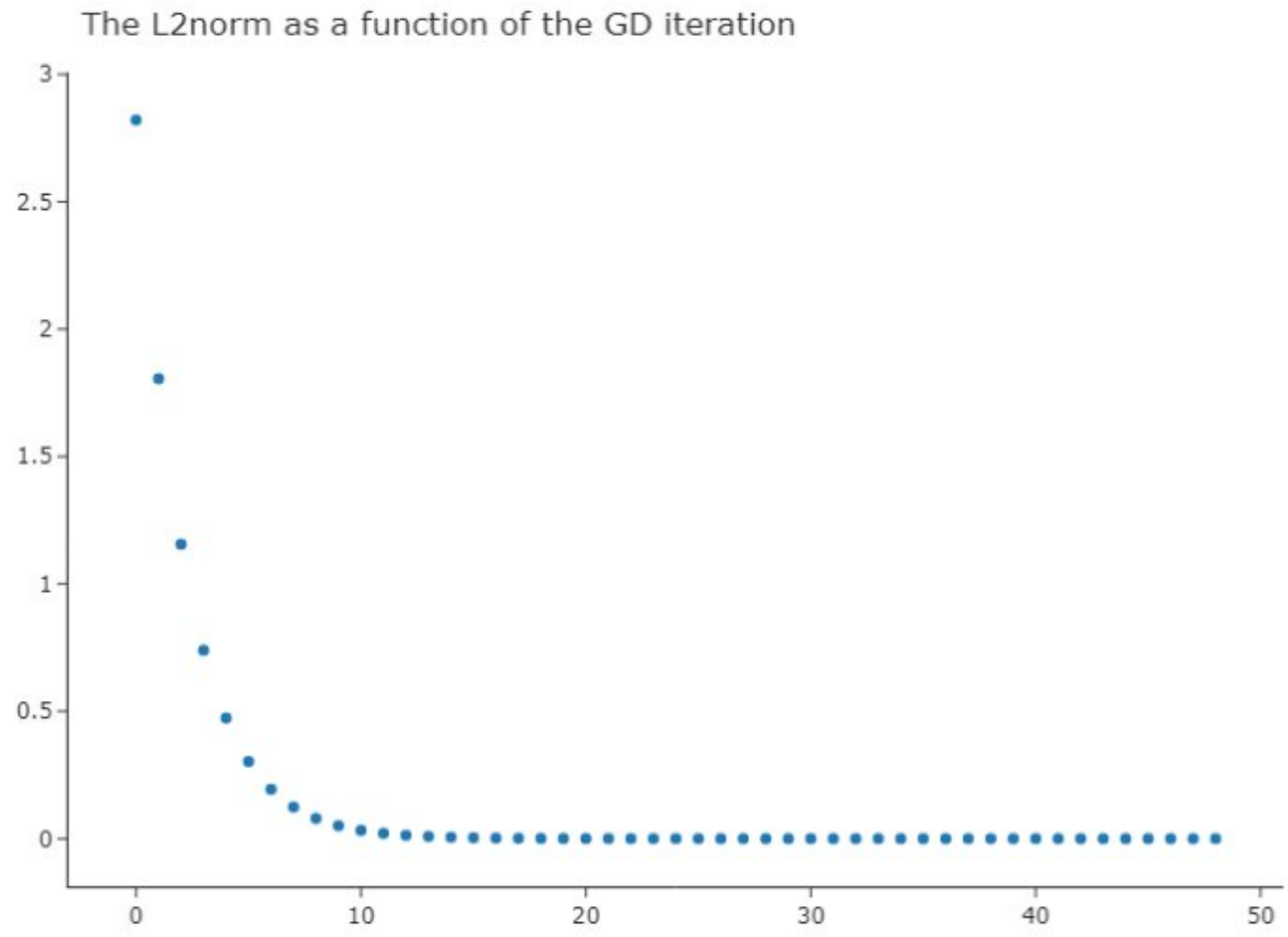
Scanned with CamScanner

(*) For $\eta = 0.1$

L1:



L2:



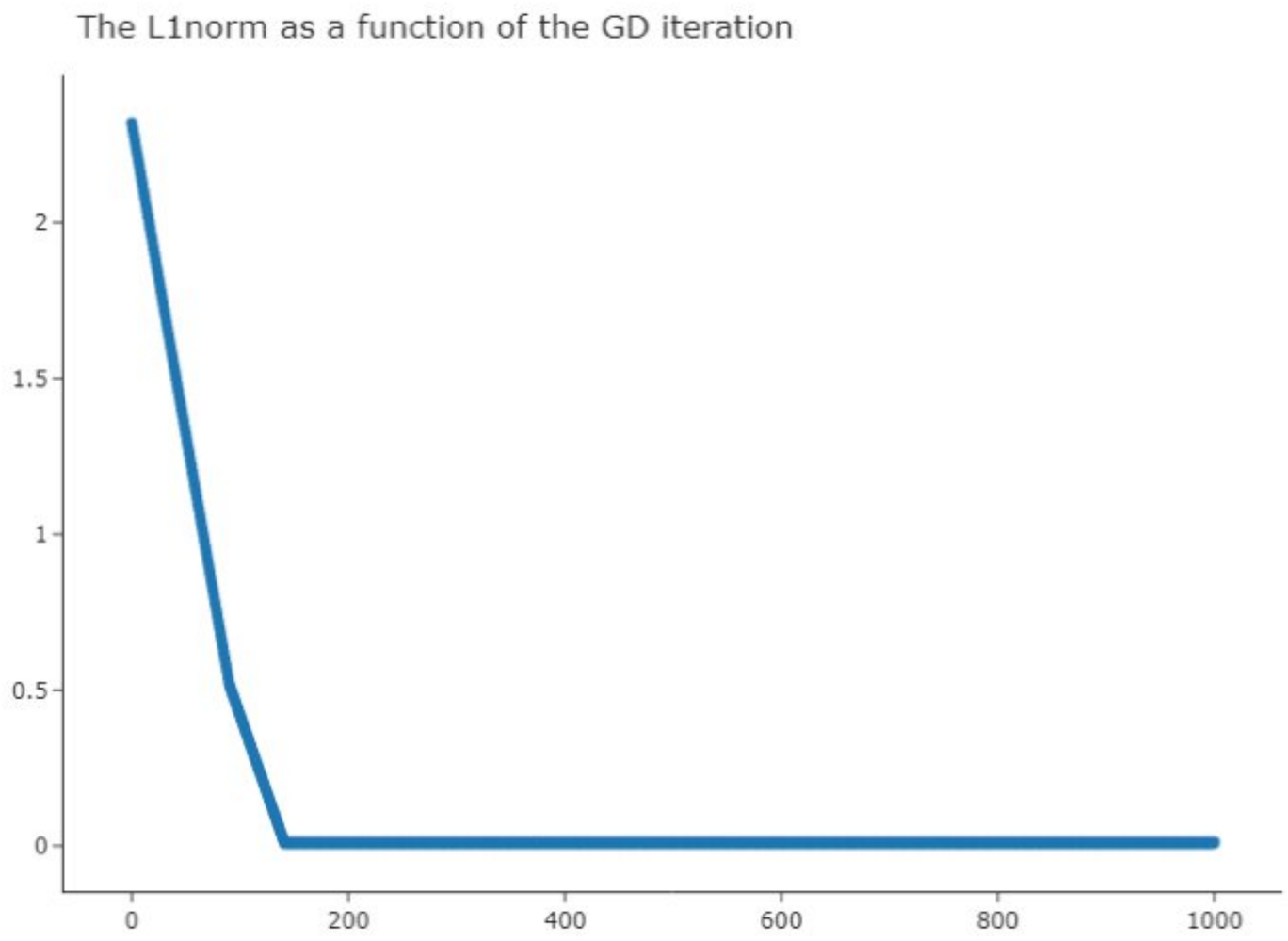
Exploitation:

For the L1 norm the learning constant is almost good because the resulting norms are small and approach zero after a small number of iterations. However, zero norm is not obtained and in the last iterations the jump is between 2 fixed points. In contrast, the L2 norm with this learning constant converges to 0 after a small number of iterations..

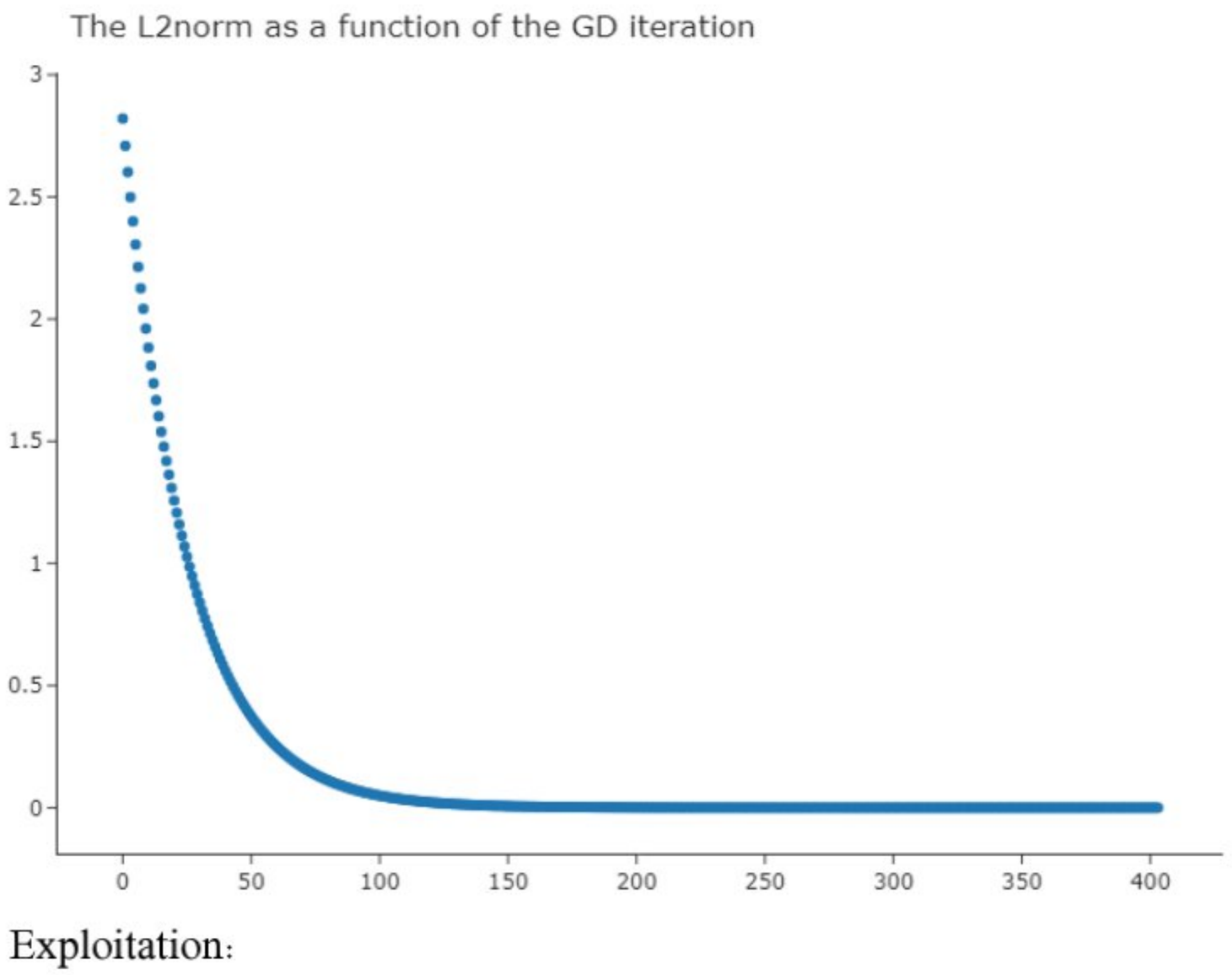
Scanned with CamScanner

(*) For $\eta = 0.01$

L1:



L2:



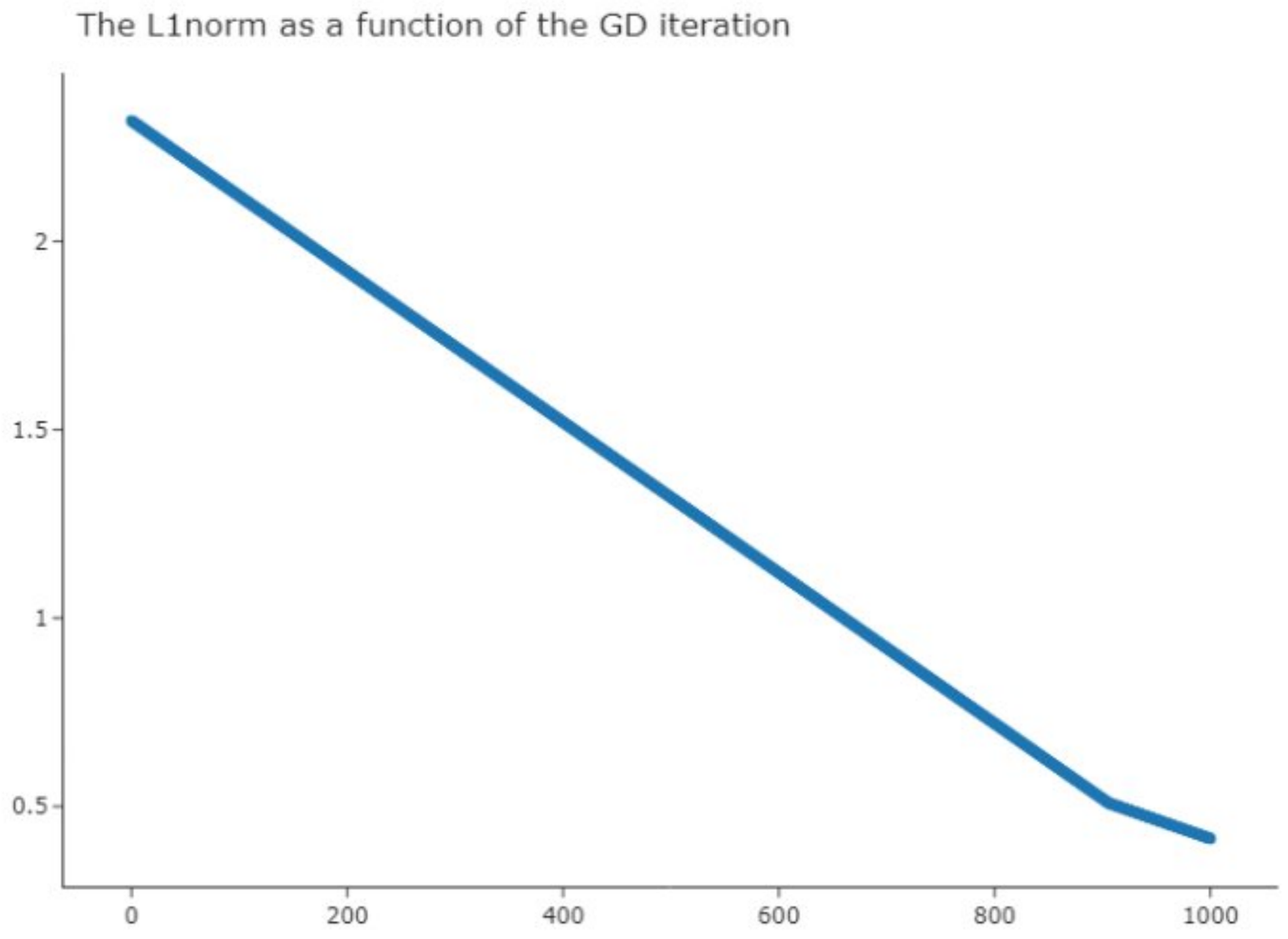
Exploitation:

For this learning constant the behavior of the 2 models is almost identical.
The norm converges to 0 in both.

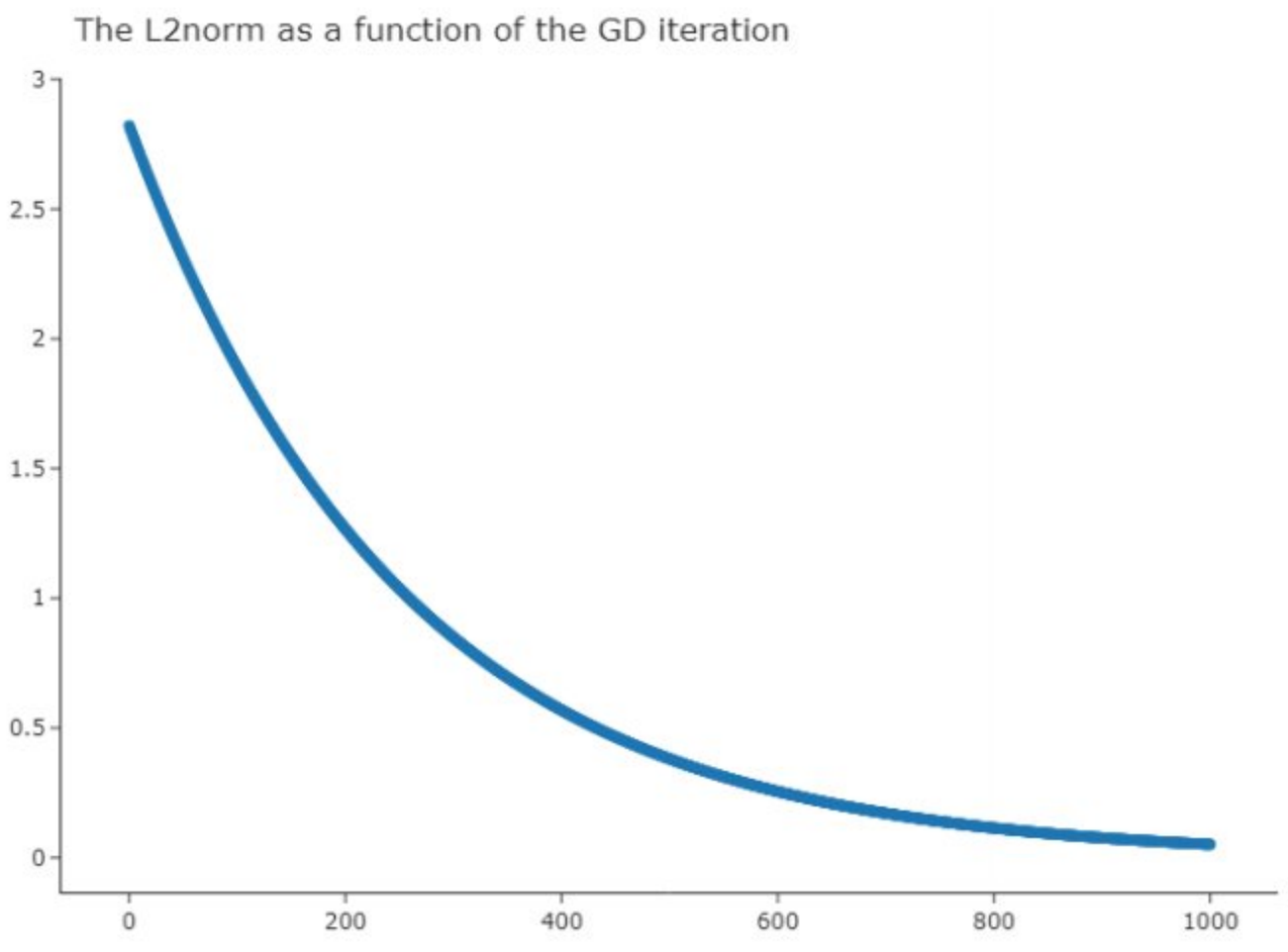
Scanned with CamScanner

(*) For $\eta = 0.001$

L1:



L2:



Exploitation:

For this learning constant, the L1 norm did not get a 0 norm because the bit of convergence to zero is low and even after 1000 iterations we didn't get the 0 norm. For L2 norm after 1000 iterations we minimized the norm and accepted the 0 norm.

Scanned with CamScanner

Q4: The lowest loss achieved when minimizing each of the modules:

```
The lowest loss achieved when minimizing norm L1 and η = 0.01 is :0.008119619553413011
The lowest loss achieved when minimizing norm L2 and η = 0.1 is :1.4029519498344255e-09
```

Exploitation:

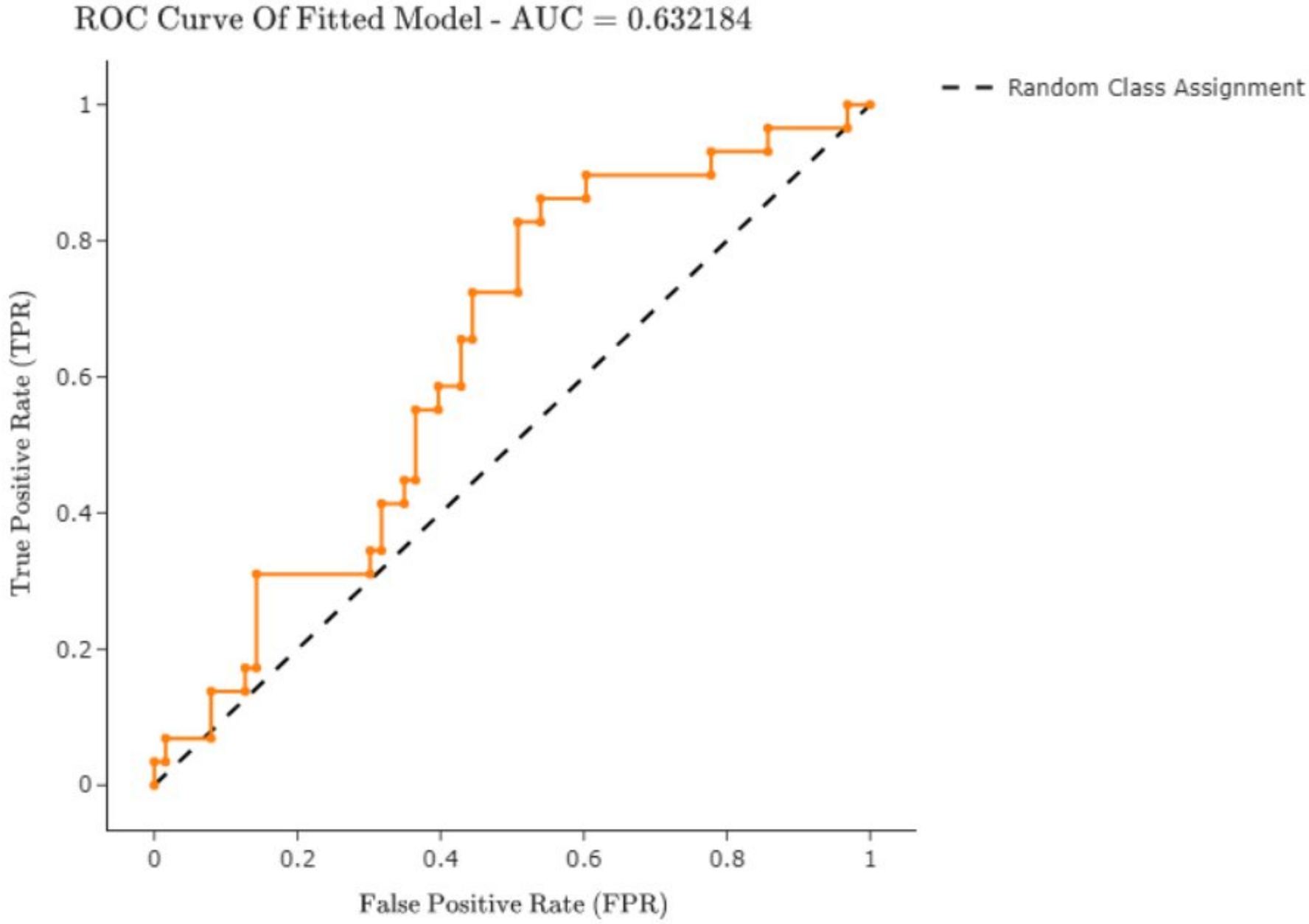
The derivatives of the norms are different and therefore different learning constants were required to reach minimal LOSS.

Part 2.1.1 Comparing Exponentially Decaying learning rates (Optional):

[Q5- Q7: Optional]

Part 2.2: Minimizing Regularized Logistic Regression:

Q8: Plot of ROC curve - logistic regression model over the data:



Scanned with CamScanner

Q9:

(*) Value of α achieves the optimal ROC:

```
Value of α achieves the optimal ROC value is: 0.0011548724510981264
```

(*) Using this value of α * the model's test error:

```
model's test error of α achieves the optimal ROC value is: 0.6847826086956522
```

Q10: ℓ1-regularized logistic regression -value of λ was selected and the model's test error:

```
Fitting l1 regularized logistic regression model:
    (*) Value of λ was selected is: 0.02
    (*)The model's test error is: 0.2826086956521739
```

Q11: ℓ2-regularized logistic regression -value of λ was selected and the model's test error:

```
Fitting l2 regularized logistic regression model:
    (*) Value of λ was selected is: 0.1
    (*)The model's test error is: 0.3260869565217391
```

Scanned with CamScanner