# Theory Questions

* 1. Let there be and as required.

Let there be .

* 1. Let there be as required.

Let there be .

* 1. We will show that is convex:

Because and for every , the second derivative is positive for every , hence is convex.

In the first article of this question, we showed that for any convex function , and , is convex.

Therefore, if we take , and set as the zero vector in we can conclude that is convex.

Because this is a composition of convex function with affine function, as in article (a).

1. We will prove that for satisfies that .

First, we will prove the following **proposition**:

**,** when

.

Prove:

If then

If then

Therefore,

Now we focus on proving that is also optimal for

Let there be .

Let us see how behaves:

Because for all .

If we take , we notice that ,

Hence

Using the hint, if we take , we get this constraint () for all .

So, we conclude that for large enough , ,

So, .

Therefore,

Where the last statements follow from the fact that is non-negative.

Because the hinge lost bounds the zero-one loss, we conclude that

And this happens only if

Let there be it’s projection on .

Because is convex, for any we have that

Therefore, because

So, we conclude that .

From taking we can see that

So,

Where in the last inequality we used that   
and this yields us that

Let there be , and as in the Theorem.

I assume that , for every , and (like what was presented in Lecture #5).

Using Jensen inequality, we obtain that

Because is convex, we get that for any

Thus,

And because

we get that

Using completing the square we get that

Therefore,

And if we again use that we get

From article (a), we obtain that

And this is the same as

So, if we plug this into the upper formula, we get that

For conclusion, we got that

So,

1. Let there be as in the question (differential, non-negative and smooth). Let us consider points in GD, where .

Because is smooth:

Therefore,

Let’s denote the partial sums sequence.

We obtain from the above that:

Let us calculate when :

And from our assumptions, this holds.

So, we can use that to obtain:

This means that is bounded from above by non-negative number:

We notice that , because

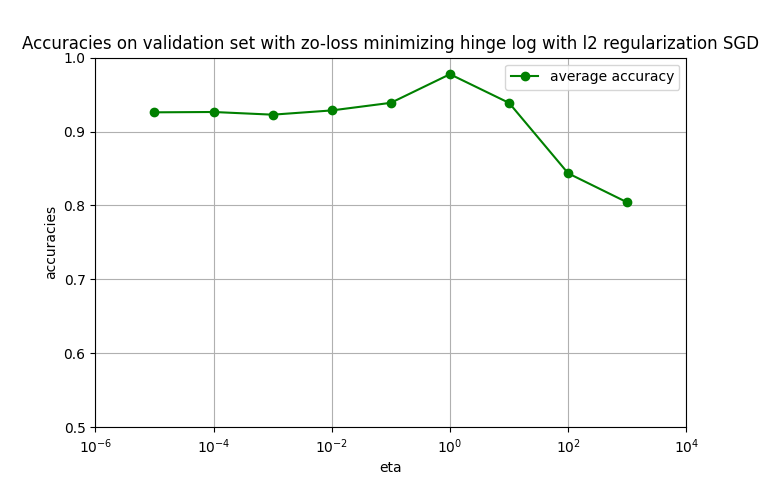
()

This means that converges because non-decreasing and bounded from above sequence converges to its supremum.

Therefore from the hint, and therefore

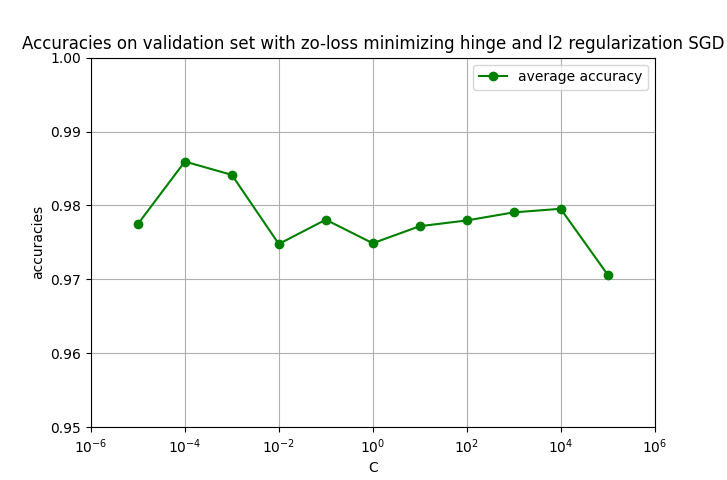
(If , then ).

# Programming Questions

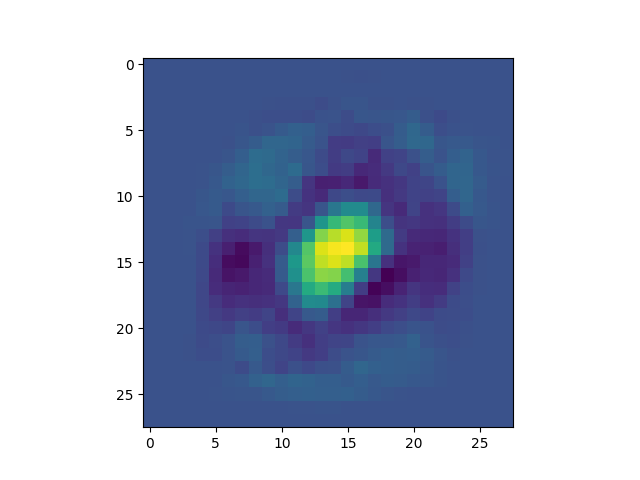


When was larger than there overflow encountered in multiplication, so we tested up to .

The best is 1.

1. 

The best is 0.0001.

1. 

We can interpretate that the “warmer” the color, the result of SGD is denser, and when the color is “cooler”, it’s sparser.

1. The accuracy on the test dataset is 99.28352098259979%.
   1. Chart, line chart

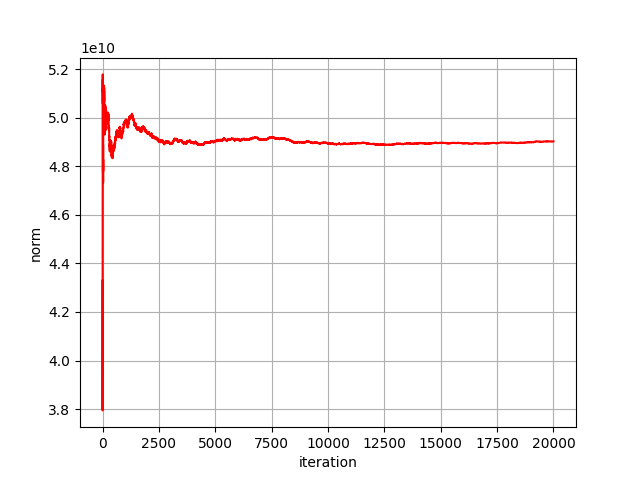
      Description automatically generated

The best is 0.1.

* 1. Chart

     Description automatically generated with medium confidence

The accuracy is 75.84442169907881%.

* 1. 

We can see that the norms sequence converges.

We know that in expectation SGD converges to the optimal point.

Probably, this number of iterations was large enough for SGD to convert for optimum, and because of that the norms sequence convergent to the norm of the optimum.