Theory Questions

1.

a. Let there be f, A and b as required.
Let there be $x, y \in \mathbb{R}^n, \lambda \in [0,1]$.

$$g(\lambda x + (1-\lambda)y) = f(A\lambda x + A(1-\lambda)y + b) =$$
$$f(\lambda Ax + \lambda b + (1-\lambda)Ay + (1-\lambda)b) = f(\lambda(Ax+b) + (1-\lambda)(Ay+b) \leq$$
$$\lambda f(Ax+b) + (1-\lambda)f(Ay+b) = \lambda g(x) + (1-\lambda)g(y)$$

b. Let there be $\{f_i\}_{i=1}^m$ as required.
Let there be $x, y \in \mathbb{R}^n, \lambda \in [0,1]$.

$$g(\lambda x + (1-\lambda)y) = \max_i f_i(\lambda x + (1-\lambda)y)$$
$$\leq \max_i \lambda f_i(x) + (1-\lambda)f_i(y)$$
$$\leq \max_i \lambda f_i(x) + \max_i (1-\lambda)f_i(y) = \lambda g(x) + (1-\lambda)g(y)$$

c. We will show that $\ell_{\log}$ is convex:

$$\frac{d\ell_{\log}}{dz} = \frac{-e^{-z}}{(1+e^{-z})ln2}$$

$$\frac{d^2\ell_{\log}}{dz^2} = \frac{e^{-z}(1+e^{-z})ln2 - (-e^{-z})ln2(-e^{-z})}{\left((1+e^{-z})ln2\right)^2}$$
$$= \frac{ln2e^{-z} + ln2e^{-2z} - ln2e^{-2z}}{\left((1+e^{-z})ln2\right)^2}$$
$$= \frac{ln2e^{-z}}{\left((1+e^{-z})ln2\right)^2}$$

Because $ln2 > 0$, and for every $z \in \mathbb{R}, e^{-z} > 0$, the second derivative $\frac{d^2\ell_{\log}}{dz^2}$ is positive for every $z \in \mathbb{R}$, hence $\ell_{\log}$ is convex.

In the first article of this question, we showed that for any convex function $f: \mathbb{R}^n \to \mathbb{R}$, $A \in \mathbb{R}^{n \times n}$, and $b \in \mathbb{R}^n$, $g(x) = f(Ax+b)$ is convex.

Therefore, if we take $y \in \mathbb{R}, \boldsymbol{x} \in \mathbb{R}^n$, and set b as the zero vector in $\mathbb{R}^n$ we can conclude that $f: \mathbb{R}^n \to \mathbb{R}$, $f(\boldsymbol{w}) = \ell_{\log}(y\boldsymbol{w} \cdot \boldsymbol{x})$ is convex.
Because this is a composition of convex function with affine function, as in article (a).

2. We will prove that for $w^*_{hinge} \in argmin_{w\in\mathbb{R}^d} \sum_{i=1}^n \ell_{hinge}(y_i \boldsymbol{w} \cdot \boldsymbol{x}_i)$ satisfies that $\forall i\ sign(\boldsymbol{w}^*_{hinge} \cdot \boldsymbol{x}_i) = y_i$.

   First, we will prove the following **proposition**:
   $\boldsymbol{for\ any\ r} \in \mathbb{R}, \boldsymbol{\ell_{zo}(r) \le \ell_{hinge}(r)}$, when $\ell_{zo}(r) = \begin{cases} 0 & r > 0, \\ 1 & r \le 0 \end{cases}$,
   $\ell_{hinge}(r) = \max\{0, 1 - r\}$.

   Prove:
   If $r > 0$, then $\ell_{hinge}(r) = \max\{0, 1-r\} \ge 0 = \ell_{zo}(r)$
   If $r \le 0$, then $\ell_{hinge}(r) = \max\{0, 1-r\} \ge 1 = \ell_{zo}(r)$
   Therefore, $for\ any\ r \in \mathbb{R}, \ell_{zo}(r) \le \ell_{hinge}(r)$.

   Now we focus on proving that $\boldsymbol{w}^*_{hinge}$ is also optimal for ℓ_{zo}.
   Let there be $i \in \{1, \dots, n\}, c \in \mathbb{R}$.
   Let us see how $c\boldsymbol{w}^*$ behaves: $y_i c\boldsymbol{w}^* \cdot \boldsymbol{x}_i = c y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i > 0$
   Because $y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i > 0$ for all i.

   If we take $c > \frac{1}{y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i}$, we notice that $y_i c\boldsymbol{w}^* \cdot \boldsymbol{x}_i = c y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i > \frac{1}{y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i} \cdot y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i = 1$,
   Hence $\ell_{hinge}(y_i c\boldsymbol{w}^* \cdot \boldsymbol{x}_i) = 0$

   Using the hint, if we take $c \to \infty$, we get this constraint ($c > \frac{1}{y_i \boldsymbol{w}^* \cdot \boldsymbol{x}_i}$) for all i.
   So, we conclude that for large enough c, $\forall i\ \ \ell_{hinge}(y_i c\boldsymbol{w}^* \cdot \boldsymbol{x}_i) = 0$,
   So, $\sum_{i=0}^n \ell_{hinge}(y_i c\boldsymbol{w}^* \cdot \boldsymbol{x}_i) = 0$.

   Therefore,

$$0 \le \sum_{i=0}^{n} \ell_{hinge}(y_i w^*_{hinge} \cdot \boldsymbol{x}_i) \le 0$$
$$\Rightarrow \sum_{i=0}^{n} \ell_{hinge}(y_i w^*_{hinge} \cdot \boldsymbol{x}_i) = 0$$
$$\Rightarrow \forall i\ \ \ell_{hinge}(y_i w^*_{hinge} \cdot \boldsymbol{x}_i) = 0$$

   Where the last statements follow from the fact that ℓ_{hinge} is non-negative.

   Because the hinge lost bounds the zero-one loss, we conclude that
$$\forall i\ \ \ell_{zo}(y_i w^*_{hinge} \cdot \boldsymbol{x}_i) = 0$$
   And this happens only if
$$\forall i\ \ sign(w^*_{hinge} \cdot \boldsymbol{x}_i) = y_i$$

3.

a.

Let there be $y \in \mathbb{R}^d, x = \prod_K(y)$ it's projection on K.

Because K is convex, for any $\lambda \in (0,1), z \in K$ we have that

$$\lambda z + (1-\lambda)x = x + \lambda(z-x) \in K$$

Therefore, because $\|x-y\|_2 = \min_{z \in K} \|y-z\|$

$$\|x-y\|_2^2 \leq \|x + \lambda(z-x) - y\|_2^2$$
$$= \|x-y\|_2^2 + 2\lambda\langle x-y, z-x\rangle + \lambda^2\|z-x\|_2^2$$

So, we conclude that $2\lambda\langle x-y, z-x\rangle \geq -\lambda^2\|z-x\|_2^2$.
From taking $\lambda \to 0$ we can see that $\langle x-y, z-x\rangle \geq 0$

So,

$$\|y-z\|_2^2 = \|z-y\|_2^2 = \|z-x+x-y\|_2^2$$
$$= \|z-x\|_2^2 + 2\langle x-y, z-x\rangle + \|x-y\|_2^2$$
$$\geq \|z-x\|_2^2$$

Where in the last inequality we used that $\langle x-y, z-x\rangle \geq 0, \|x-y\|_2^2 \geq 0$.
and this yields us that $\|y-z\|_2 \geq \|x-z\|_2$

b.

Let there be $x^* \in argmin_{x \in K} f(x)$, $\eta_t = \frac{\epsilon}{G^2}$ and $T = \frac{B^2G^2}{\epsilon^2}$ as in the Theorem.

I assume that $\|x^*\|_2^2 \leq B$, $\|\nabla f(x_t)\|_2^2 \leq G$ for every $x \in K$, and $x_1 = \mathbf{0}$ (like what was presented in Lecture #5).

Using Jensen inequality, we obtain that

$$f(\bar{x}) - f(x^*) = f\left(\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T} f(x_t) - f(x^*)$$
$$= \frac{1}{T}\sum_{t=1}^{T}(f(x_t) - f(x^*))$$

Because f is convex, we get that for any t,

$$f(x_t) - f(x^*) \leq \langle \nabla f(x_t), x_t - x^*\rangle$$

Thus,

$$f(\bar{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x^*\rangle$$

And because $y_{t+1} = x_t - \eta_t \nabla f(x_t)$,
we get that

$$f(\bar{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\langle \nabla f(x_t), x_t - x^* \rangle \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta_t}\langle \mathrm{x_t} - y_{t+1}, x_t - x^* \rangle$$

Using completing the square we get that

$$(a-b)(a-c) = \frac{(a-b)^2 + (a-c)^2 - (b-c)^2}{2}$$

Therefore,

$$\frac{1}{\eta_t}\langle \mathrm{x_t} - y_{t+1}, x_t - x^* \rangle$$

$$= \frac{1}{2\eta_t}(\|\mathrm{x_t} - y_{t+1}\|_2^2 + \|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2)$$

And if we again use that $y_{t+1} = x_t - \eta_t \nabla f(x_t)$, we get

$$= \frac{1}{2\eta_t}(\|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2$$

From article (a), we obtain that

$$\|x_{t+1} - x^*\|_2 \leq \|y_{t+1} - x^*\|_2$$

And this is the same as

$$-\|x_{t+1} - x^*\|_2 \geq -\|y_{t+1} - x^*\|_2$$

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

$$\frac{1}{2\eta_t}(\|x_t - x^*\|_2^2 - \|y_{t+1} - x^*\|_2^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2$$

$$\leq \frac{1}{2\eta_t}(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2$$

For conclusion, we got that

$$\frac{1}{\eta_t}\langle \mathrm{x_t} - y_{t+1}, x_t - x^* \rangle \leq \frac{1}{2\eta_t}(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2$$

So,

$$f(\bar{x}) - f(x^*) \leq \frac{1}{T}\sum_{t=1}^{T}\frac{1}{\eta_t}\langle \mathrm{x_t} - y_{t+1}, x_t - x^* \rangle$$

$$\leq \frac{1}{T}\sum_{t=1}^{T}\left(\frac{1}{2\eta_t}(\|x_t - x^*\|_2^2 - \|x_{t+1} - x^*\|_2^2) + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2\right)$$

$$= \frac{1}{2T\eta_t}(\|x_1 - x^*\|_2^2 - \|x_{T+1} - x^*\|_2^2) + \frac{\eta_t}{2T}\|\nabla f(x_t)\|_2^2$$

$$\leq \frac{1}{2T\eta_t}\|x_1 - x^*\|_2^2 + \frac{\eta_t}{2T}\|\nabla f(x_t)\|_2^2 \leq \frac{1}{2T\eta_t}\|x^*\|_2^2 + \frac{\eta_t}{2}\|\nabla f(x_t)\|_2^2$$

$$\leq \frac{B^2}{2T\eta_t} + \frac{\eta_t}{2}G^2 = \frac{B^2}{2}\frac{G^2\epsilon^2}{\epsilon B^2 G^2} + \frac{\epsilon}{2G^2}G^2 = \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$$

4. Let there be $f: \mathbb{R}^n \to \mathbb{R}$ as in the question (differential, non-negative and β smooth). Let us consider x_t, x_{t+1} points in GD, where $\eta < \frac{2}{\beta}$.

   Because f is β smooth:

$$f(x_{t+1}) \leq f(x_t) + \nabla f(x_t)^T(x_{t+1} - x_t) + \frac{\beta}{2}\|x_{t+1} - x_t\|^2$$
$$= f(x_t) - \eta \nabla f(x_t)^T \nabla f(x_t) + \frac{\beta \eta^2}{2}\|\nabla f(x_t)\|^2$$
$$= f(x_t) - \|\nabla f(x_t)\|^2 \left(\eta - \frac{\beta}{2}\eta^2\right)$$

   Therefore,

$$\|\nabla f(x_t)\|^2 \leq \left(\eta - \frac{\beta}{2}\eta^2\right)^{-1} \left(f(x_t) - f(x_{t+1})\right)$$

   Let's denote $S_n = \sum_{t=1}^{n} \|\nabla f(x_t)\|^2$ the partial sums sequence.
   We obtain from the above that:

$$S_n \leq \left(\eta - \frac{\beta}{2}\eta^2\right)^{-1} \left(f(x_1) - f(x_{n+1})\right)$$

   Let us calculate when $\eta - \frac{\beta}{2}\eta^2 > 0$:

$$\eta - \frac{\beta}{2}\eta^2 > 0 \Leftrightarrow \eta\left(1 - \frac{\beta}{2}\eta\right) > 0 \Leftrightarrow 0 < \eta < \frac{2}{\beta}$$

   And from our assumptions, this holds.
   So, we can use that $\forall x \in \mathbb{R}^n,\ f(x) \geq 0$ to obtain:

$$S_n \leq \left(\eta - \frac{\beta}{2}\eta^2\right)^{-1} f(x_1)$$

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

$$\left(\eta - \frac{\beta}{2}\eta^2\right)^{-1} f(x_1)$$

   We notice that $S_{n+1} \geq S_n$, because
   $S_{n+1} = S_n + \|\nabla f(x_{n+1})\|^2 \geq S_n$ $(\forall t \in \mathbb{N}\ \|\nabla f(x_{n+1})\|^2 \geq 0)$

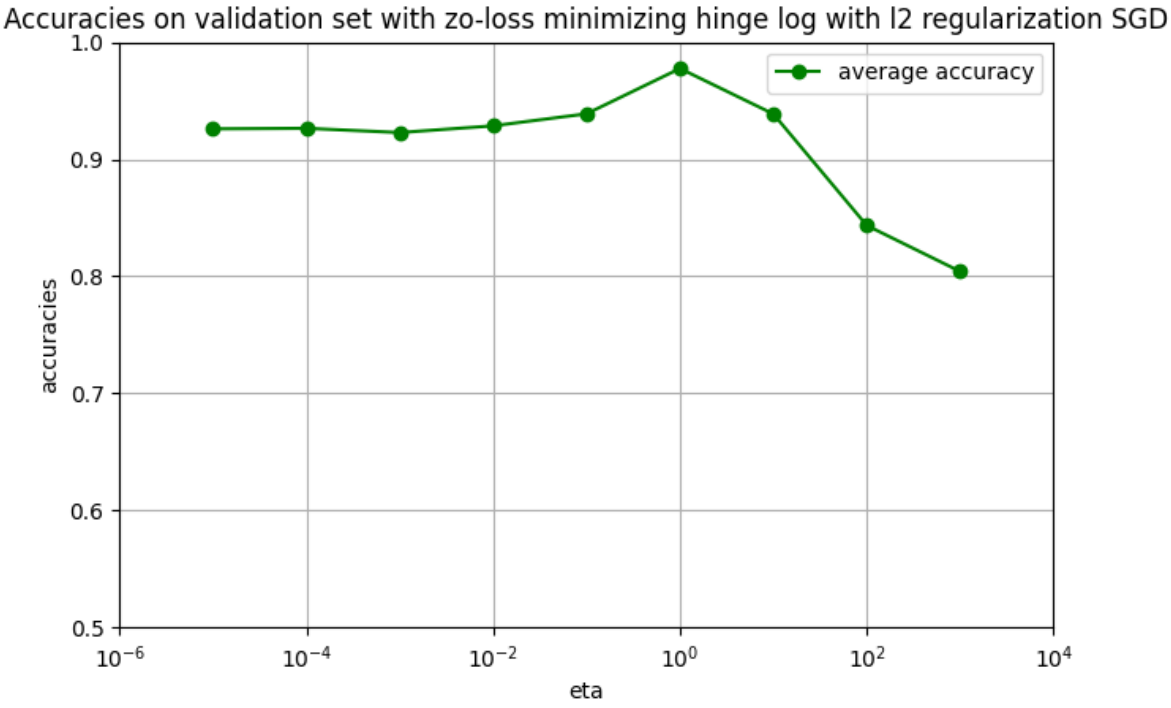
   This means that S_n converges because non-decreasing and bounded from above sequence converges to its supremum.

   Therefore $\lim_{t\to\infty} \|\nabla f(x_t)\|^2 = 0$ from the hint, and therefore $\lim_{t\to\infty} \|\nabla f(x_t)\| = 0$
   (If $\lim_{n\to\infty} a_n^2 = 0$, then $\lim_{n\to\infty} a_n = 0$).

Programming Questions

1.

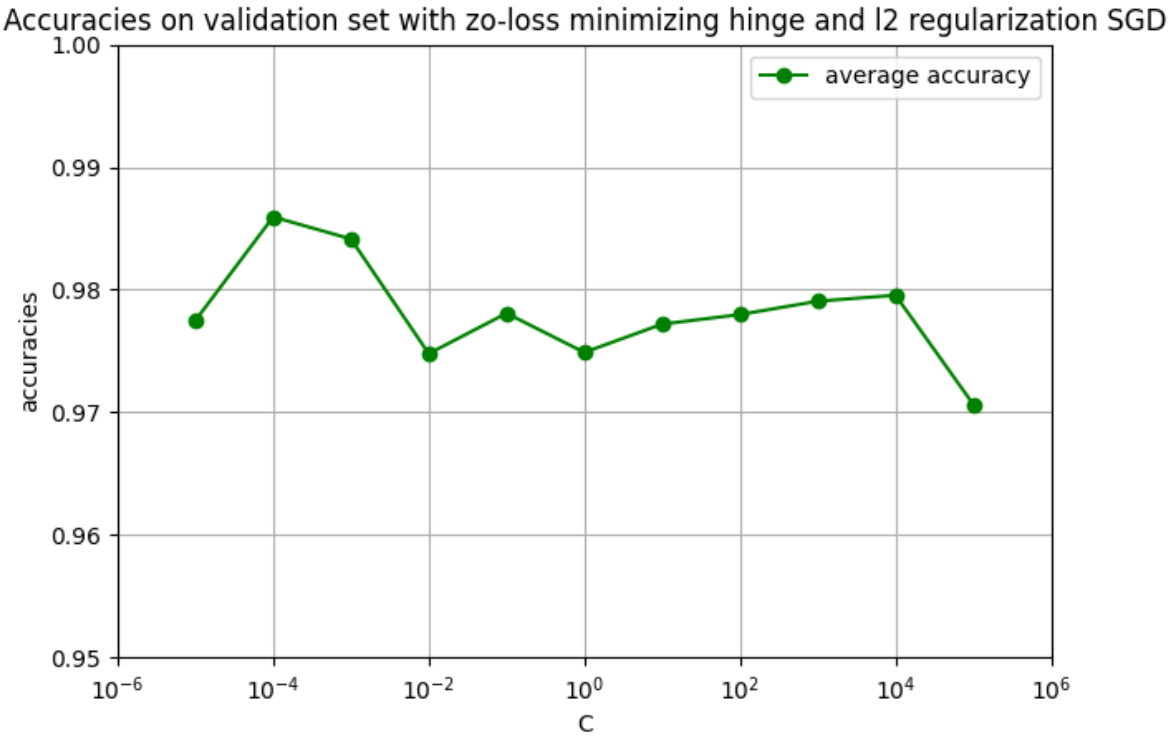
a.



When η was larger than 1000 there overflow encountered in multiplication, so we tested up to $\eta = 1000$.

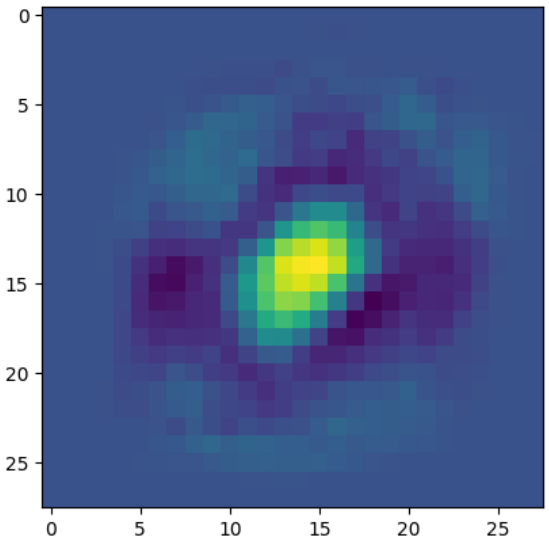
The best η is 1.

b.



The best C is 0.0001.

c.

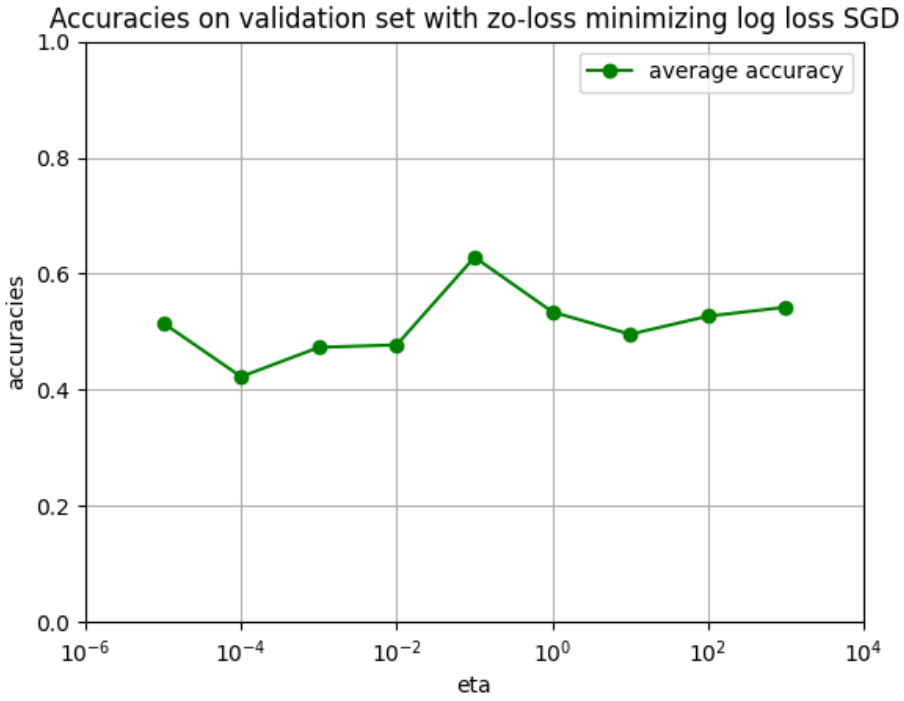


We can interpretate that the “warmer” the color, the result of SGD is denser, and when the color is “cooler”, it’s sparser.

d. The accuracy on the test dataset is 99.28352098259979%.

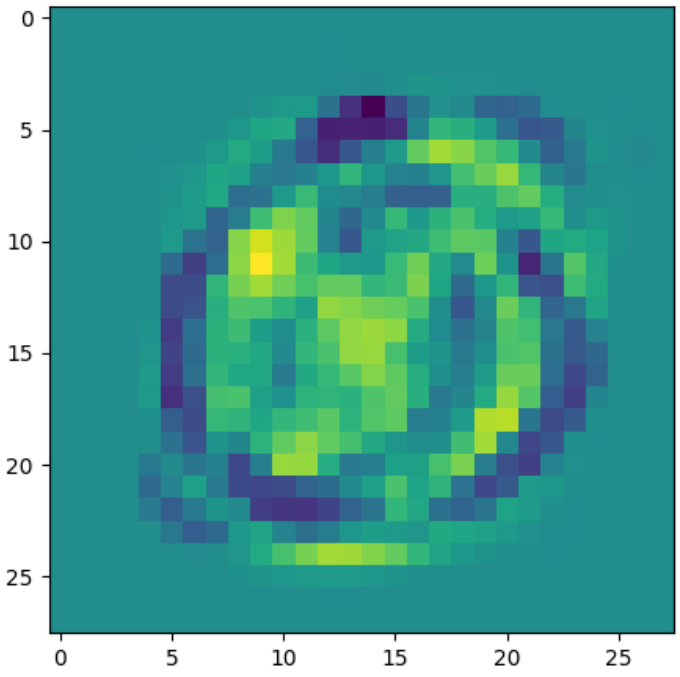
2.

a.



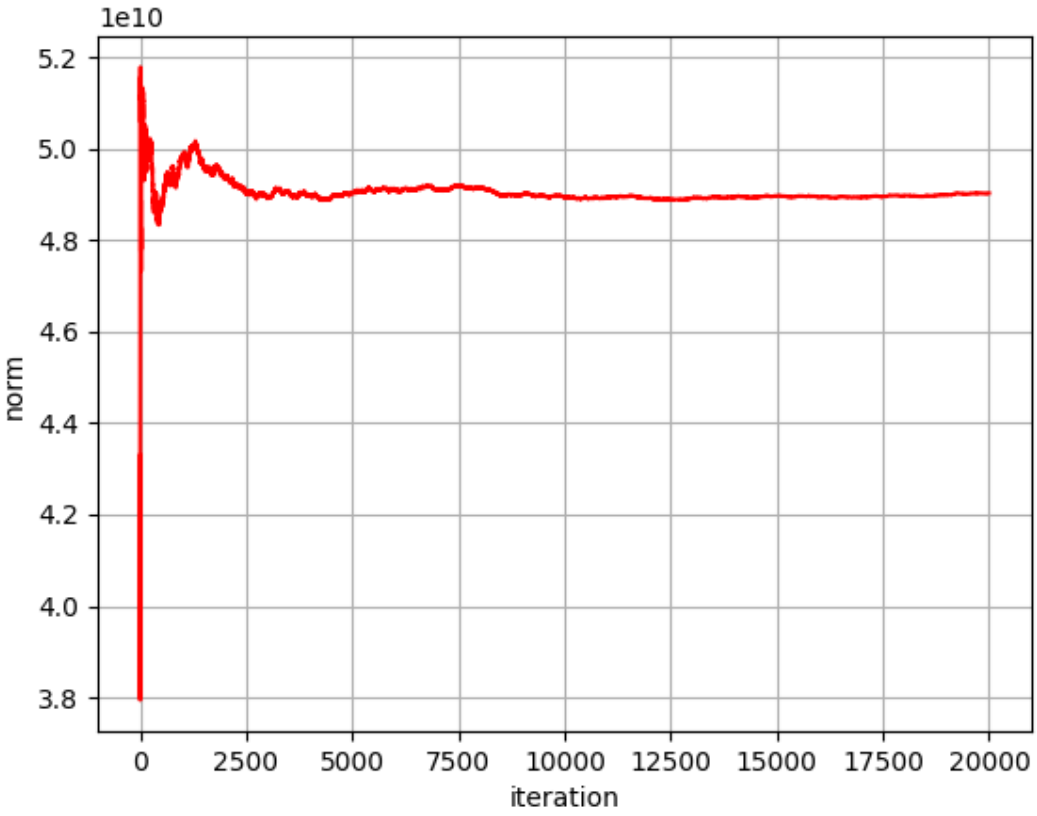
The best η is 0.1.

b.



The accuracy is 75.84442169907881%.

c.



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.