Introduction to Machine Learning: Ex03

Theoretical Questions

1.

(15 points) Step-size Perceptron. Consider the modification of Perceptron algorithm with the following update rule:

$$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t + \eta_t y_t \mathbf{x}_t$$

whenever $\hat{y}_t \neq y_t$ ($\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t$ otherwise). Assume that data is separable with margin $\gamma > 0$ and that $\|\mathbf{x}_t\| = 1$ for all t. For simplicity assume that the algorithm makes M mistakes at the first M rounds, afterwhich it has no mistakes. For $\eta_t = \frac{1}{\sqrt{t}}$, show that the number of mistakes step-size Perceptron makes is at most $\frac{4}{\gamma^2}\log^2(\frac{1}{\gamma})$. (Hint: use the fact that if $x \leq a\log(x)$ then $x \leq 2a\log(a)$).

It is given that $w_{t+1} = w_t + \frac{1}{\sqrt{t}} y_t x_t$ and $\forall_t. \|x_t\| = 1$ and $\|w_0\| = 0$, $\|w^*\| = 1$:

$$w_{t+1} \cdot w^* = \left(w_t + \frac{1}{\sqrt{t}} y_t x_t\right) w^* = w_t w^* + \frac{1}{\sqrt{t}} y_t x_t w^*$$

Since the separable margin is γ *we get* $\frac{\gamma}{\sqrt{M}} \leq \frac{1}{\sqrt{t}} y_t x_t w^*$ *so it can be shown that*

$$\gamma\sqrt{M} \leq w_{M+1} w^* \leq \|w_{M+1}\| \cdot \|w^*\| \leq \|w_{M+1}\| \to \boldsymbol{M \leq \frac{1}{\gamma^2}\|w_{M+1}\|}$$

So we will find an upper bound for $\|w_{M+1}\| \leq 4\log^2\left(\frac{1}{\gamma}\right)$

Let us remind that $\forall t \neq 0. \|x_t\| = 1$ *and* $y_t = \pm 1$ *so* $\|x_t y_t\| = 1$

$$\|w_{M+1}\|^2 = \sum_{t=1}^{M} \|w_{t+1}\|^2 - \|w_t\|^2 = \sum_{t=1}^{M} \|w_t\|^2 + \frac{2}{\sqrt{t}}\|w_t\| \cdot \|y_t x_t\| + \frac{1}{t}\|y_t x_t\| - \|w_t\|^2 = \sum_{t=1}^{M} \frac{2}{\sqrt{t}}\|w_t\| + \frac{1}{t}$$

Since we have M mistakes on the first rounds we get that

$$\forall t \leq M. \|w_t\| = \|w_0\| + \sum_{i=1}^{t} \frac{1}{i}\|y_t x_t\| = \sum_{i=1}^{t} \frac{1}{t} < \ln(t)$$

$$\|w_{M+1}\|^2 < \sum_{t=1}^{M} \frac{2}{\sqrt{t}}\ln(t) < \sum_{t=1}^{M} \frac{2}{\sqrt{t}}\ln(t) < 2M \cdot \ln(M)$$

Adding the constraint $M > 2$ *we can state that* $\ln(M) \leq \ln^4(M)$ *so*

$$\|w_{M+1}\| < \sqrt{2M} \cdot \ln^2 M$$

We say in class that $M = \frac{1}{\gamma}$ *so* $\|w_{M+1}\|^2 < \sqrt{2M+1} \cdot \ln^2\left(\frac{1}{\gamma}\right)$ *so all that is left to show is the numeric value.*

$$M \leq \frac{4}{\gamma^2}\ln^2\left(\frac{1}{\gamma}\right)$$

2.

(15 points) Convex functions.

(a) Let $f : \mathbb{R}^n \to \mathbb{R}$ a convex function, $A \in \mathbb{R}^{n\times n}$ and $b \in \mathbb{R}^n$. Show that, $g(\mathbf{x}) = f(A\mathbf{x}+b)$ is convex.

(b) Consider m convex functions $f_1(\mathbf{x}), \ldots, f_m(\mathbf{x})$, where $f_i : \mathbb{R}^d \to \mathbb{R}$. Now define a new function $g(\mathbf{x}) = \max_i f_i(\mathbf{x})$. Prove that $g(\mathbf{x})$ is a convex function. (Note that from (a) and (b) you can conclude that the hinge loss over linear classifiers is convex.)

(c) Let $\ell_{log} : \mathbb{R} \to \mathbb{R}$ be the log loss, defined by

$$\ell_{\log}(z) = \log_2\left(1+e^{-z}\right)$$

Show that ℓ_{log} is convex, and conclude that the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(\mathbf{w}) = \ell_{log}(y\mathbf{w}\cdot\mathbf{x})$ is convex with respect to $\mathbf{w}$.

a. Let there be $v \in \mathbb{R}$ and $v_1, v_2 \in \mathbb{R}^n, \beta \in [0,1]$ such that $v = \beta v_1 + (1-\beta)v_2$ *WLOG*.

$$g(v) = f(Av+b) = f(\beta Av_1 + (1-\beta)Au_2 + b)$$

$b = \beta b + (1-\beta)b$ we will get

$$g(v) = f(\beta Av_1 + (1-\beta)Av_2 + \beta b + (1-\beta)b) = f(\beta(Av_1+b) + (1-\beta)(Av_2+b))$$

We can rewrite $Av_1 + b$ as $\widetilde{v_1}$ and $Av_2 + b$ as $\widetilde{v_2}$ so:

$$g(v) = f(\beta\widetilde{v_1} + (1-\beta)\widetilde{v_2})$$

Since $f: \mathbb{R}^n \to \mathbb{R}$ is convex: $\forall x_1, x_2 \in \mathbb{R}^n, \forall\beta \in [0,1]$: $\boldsymbol{f(\beta x_1 + (1-\beta)x_2) \leq \beta f(x_1) + (1-\beta)f(x_2)}$

Applying this definition of $f(\cdot)$ on the $g(v)$ statement:

$$g(v) = g(\beta v_1 + (1-\beta)v_2) \leq f(\beta\widetilde{v_1} + (1-\beta)\widetilde{v_2}) \leq \beta f(\widetilde{v_1}) + (1-\beta)f(\widetilde{v_2})$$

And since $f(\widetilde{v_i})$ can be rewritten as $f(Av_i + b) = g(v_i)$ we can rewrite the inequality as:

$$\boldsymbol{g(\beta v_1 + (1-\beta)v_2) \leq \beta g(v_1) + (1-\beta)g(v_2)}$$

After rewriting the inequality, we received the definition for a convex function.

b.

Let there be $x_1, x_2 \in \mathbb{R}^d$, $\beta \in [0,1]$ and $f_i \in \{f_1, \cdots, f_m\}$ so we know by definition

$$\boldsymbol{f_i(\beta x_1 + (1-\beta)x_2) \leq \beta f_i(x_1) + (1-\beta)f_i(x_2)}$$

Since $\beta \geq 0$: $\max\limits_j(\beta \cdot f_i(x)) = \beta \max\limits_j(f_i(x))$ and the same with $(1-\beta)$

Denote $g(x) = \max\limits_i(f_i(x))$:

$$g(\beta x_1 + (1-\beta)x_2) = \max_i(f_i(\beta x_1 + (1-\beta)x_2)) \leq \max_i(\beta f(x_1) + (1-\beta)f(x_2))$$

$$g(\beta x_1 + (1-\beta)x_2) \leq \beta \max_i(f(x_1)) + (1-\beta)\max_i(f(x_2))$$

Since $\beta \max\limits_i(f(x_1)) + (1-\beta)\max\limits_i(f(x_2)) \equiv \beta g(x_1) + (1-\beta)g(x_2)$ we (once again) reached the convex definition:

$$\boldsymbol{(\beta v_1 + (1-\beta)v_2) \leq \beta g(v_1) + (1-\beta)g(v_2)}$$

c.

Let $\ell: \mathbb{R} \to \mathbb{R}$ be $\ell(z) = log_2(1+e^{-z})$

we will show that $f(x) = \log_2(e^{-x})$ is convex:

Remark that $\log_2(x) = \frac{\ln(x)}{\ln(2)}$ so we will mark $\varphi = \frac{1}{\ln(2)}$ and $f(x) = \varphi\ln(e^{-x}) = -\varphi x$:

$$f(\beta x_1 + (1-\beta)x_2) = -\varphi(\beta x_1 + (1-\beta)x_2) = \beta(-\varphi x_1) + (1-\beta)(-\varphi x_2) = \beta f(x_1) + (1-\beta)f(x_2)$$

$\forall\beta \in (0,1), x_1, x_2 . \log_2(\beta x_1 + (1-\beta)x_2) = \beta f(x_1)$

So $f(x) = \log_2(e^{-x})$ is convex, according to section A if $f(x)$ is convex so do $f(ax+b)$

$\boldsymbol{f(ax+b) = \log_2(ae^{-x}+b)}$ **is convex and specifically** $\boldsymbol{\ell(\mathrm{x}) = \log(1+e^{-x})}$ **is convex** $\boldsymbol{(a=1, b=1)}$

3.

(20 points) GD with projection. In the context of convex optimization, sometimes we would like to limit our solution to a convex set $\mathcal{K} \subseteq \mathbb{R}^d$; that is,

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{K}$$

for a convex function f and a convex set $\mathcal{K}$. In this scenario, each step in the gradient descent algorithm might result in a point outside $\mathcal{K}$. Therefore, we add an additional projection step. The projection operator finds the closest point in the set, i.e.:

$$\Pi_{\mathcal{K}}(\mathbf{y}) := \arg\min_{\mathbf{x}\in\mathcal{K}} \|\mathbf{x}-\mathbf{y}\|_2$$

A modified iteration in the *gradient descent with projection* therefore consists of:

$$\mathbf{y}_{t+1} = \mathbf{x}_t - \eta_t \nabla f(\mathbf{x}_t)$$
$$\mathbf{x}_{t+1} = \Pi_{\mathcal{K}}(\mathbf{y}_{t+1})$$

(a) Let $\mathbf{y} \in \mathbb{R}^d$ and $\mathbf{x} = \Pi_{\mathcal{K}}(\mathbf{y})$. Prove that for any $\mathbf{z} \in \mathcal{K}$, we have $\|\mathbf{y}-\mathbf{z}\|_2 \geq \|\mathbf{x}-\mathbf{z}\|_2$. (**Guidance:** use the projection definition and the fact that for any $\lambda \in (0,1)$, $(1-\lambda)\mathbf{x} + \lambda\mathbf{z} \in \mathcal{K}$ to show that $\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{z}\rangle \leq \lambda\|\mathbf{z}-\mathbf{x}\|_2^2$ for any $\lambda \in (0,1)$. Conclude that $\langle \mathbf{x}-\mathbf{y}, \mathbf{x}-\mathbf{z}\rangle \leq 0$. Use that to show the claim in the question.)

(b) Prove that the convergence theorem for GD still holds.

a.

By definition x is the closest point in $\mathcal{K}$ to y: $x = \Pi_{\mathcal{K}}(y) = argmin_{x\in\mathcal{K}}\|x-y\|_2$
So $\forall z \in \mathcal{K}.\ \|x-y\|_2 \leq \|z-y\|_2$.
If $y \in \mathcal{K}$ we get $x = y$ so *it is pretty trivial that* $z - x = z - y$ *so* $\|z-x\|_2 = \|z-y\|_2 \leq \|z-y\|_2$
If $y \notin \mathcal{K}$ we would assume *that the condition isn't satisfy in order to contradict It* $\|z-y\|_2 > \|z-x\|_2$:
forming a triangle x, y, z will result edges with lengths: $(l_1 = \|y-x\|_2, l_2 = \|x-z\|_2, l_3 = \|z-y\|_2)$
this means that the angle at vertex y is larger then $\frac{\pi}{2}$
Assuming $\|y-z\|_2 > \|x-z\|_2$ and *the triangle above we know:* $l_3 < l_2 \leq l_1$
Using the fact that $\mathcal{K}$ is convex set and $x, z \in \mathcal{K}$ we know that all the point on the edge between them (l_2) is also in $\mathcal{K}$.the shortest distance from y to l_2 will be the height which meet l_2 in a point we name $h \in \mathcal{K}$. This can be written mathematically:

$$\forall p \in l_2 = \{x + \lambda(z-x) |\ \lambda \in (0,1)\}.\ p \neq h \rightarrow \|h-y\|_2 < \|p-y\|_2$$

Since $x \in l_2$ and according to the definition of x we get that $x = h$. Since the height is actually an edge we got a contradiction to angle at y! So our assumption was wrong hence $\|\boldsymbol{z}-\boldsymbol{x}\|_2 \leq \|\boldsymbol{z}-\boldsymbol{y}\|_2$

b. We will use the proof for GS convergence as a foundation

Reminder: $\bar{\boldsymbol{w}} = \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w_t}$, G s.t $\forall x.\ |\nabla f(x)| < G$, w^* optimal vector in $\mathcal{K}$, B s.t $\|w_{opt}\| < B$ & $\varepsilon > 0$

We will not redo the 3 steps on the beginning since they are based on **Jensen's Inequality**, some algebraic steps and the **convexity** of f

Since w_t is defined different from original we will replace the 4th step:

Let us remember that $\forall \boldsymbol{x}, \boldsymbol{y} \in \boldsymbol{D^n}.\ \|\boldsymbol{x}\|_2^2 - \|\boldsymbol{x}-\boldsymbol{\eta y}\|_2^2 = \boldsymbol{2\eta xy} - \boldsymbol{\eta^2}\|\boldsymbol{y}\|_2^2$

$$\boldsymbol{xy} = \frac{\|x\|_2^2 - \|x-\eta y\|_2^2}{2\eta} + \frac{\eta}{2}\|y\|_2^2$$

So by choosing $x = (w_t - w^*)$ and $y = \nabla f(w_t)$ we get that:

$$\frac{1}{T}\sum_{t=1}^{T}\nabla f(\boldsymbol{w_t})\cdot(\boldsymbol{w_t}-\boldsymbol{w}^*) = \frac{1}{T}\sum_{t=1}^{T}\frac{\|\boldsymbol{w_t}-\boldsymbol{w}^*\|^2 - \|\boldsymbol{w_t}-\boldsymbol{w}^*-\eta\nabla f(\boldsymbol{w_t})\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(\boldsymbol{w_t})\|^2$$

Since $w_{t+1} = \Pi_{\mathcal{K}}(w_t - \eta\nabla f(w_t))$ and $w^* \in \mathcal{K}$ we know (according to previous section):

$$x = w_{t+1}, y = w_t - \eta\nabla f(w_t), z = w^* \rightarrow \|\boldsymbol{w_t}-\eta\nabla f(\boldsymbol{w_t})-\boldsymbol{w}^*\| \geq \|\boldsymbol{w_{t+1}}-\boldsymbol{w}^*\|$$

$$\frac{1}{T}\sum_{t=1}^{T}\frac{\|\boldsymbol{w_t}-\boldsymbol{w}^*\|^2 - \|\boldsymbol{w_t}-\boldsymbol{w}^*-\eta\nabla f(\boldsymbol{w_t})\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(\boldsymbol{w_t})\|^2 \leq \frac{1}{T}\sum_{t=1}^{T}\frac{\|\boldsymbol{w_t}-\boldsymbol{w}^*\|^2 - \|\boldsymbol{w_{t+1}}-\boldsymbol{w}^*\|^2}{2\eta} + \frac{\eta}{2}\|\nabla f(\boldsymbol{w_t})\|^2$$

The 5th and 6th steps in the are still valid.

(If I had more time, I would have proven it from scratch. since we saw it , I found it appropriate to prove only the modification needed.)

4.

(15 points) Gradient Descent on Smooth Functions. We say that a continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is β-smooth if for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^T(\mathbf{y} - \mathbf{x}) + \frac{\beta}{2}\|\mathbf{x} - \mathbf{y}\|^2$$

In words, β-smoothness of a function f means that at every point $\mathbf{x}$, f is upper bounded by a qaudratic function which coincides with f at $\mathbf{x}$.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth and non-negative function (i.e., $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$). Consider the (non-stochastic) gradient descent algorithm applied on f with constant step size $\eta > 0$:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\nabla f(\mathbf{x}_t)$$

Assume that gradient descent is initialized at some point $\mathbf{x}_0$. Show that if $\eta < \frac{2}{\beta}$ then

$$\lim_{t\to\infty}\|\nabla f(\mathbf{x}_t)\| = 0$$

(Hint: Use the smoothness definition with points $\mathbf{x}_{t+1}$ and $\mathbf{x}_t$ to show that $\sum_{t=0}^{\infty}\|\nabla f(\mathbf{x}_t)\|^2 < \infty$ and recall that for a sequence $a_n \geq 0$, $\sum_{n=1}^{\infty} a_n < \infty$ implies $\lim_{n\to\infty} a_n = 0$. Note that f is not assumed to be convex!)

In order to show that $\lim_{t\to\infty}\|\nabla f(x_t)\| = 0$ for $\eta \in \left(0, \frac{2}{\beta}\right)$ when $f(\cdot)$ is β-smooth we will use the fact that

$$x_{t+1} = x_t - \eta\nabla f(x_t) \to (x_{t+1} - x_t) = -\eta\nabla f(x_t)$$

Since f is β-smooth:

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

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

Denote $\gamma = \eta\left(\frac{\beta}{2}\eta - 1\right)$ so $\forall\eta < \frac{2}{\beta}. \left(\frac{\beta}{2}\eta - 1\right) < 0 \to \gamma < 0$ so diving both sections of the blue inequality by γ will flip the direction of it (remember γ is negative so $-\gamma = |\gamma|$:

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

Applying an infinite sum of both sections will result on the right section a telescopic series:

$$\sum_{t=0}^{\infty}\|\nabla f(x_t)\|^2 \leq \frac{1}{\gamma}\sum_{t=0}^{\infty}\left(f(x_{t+1}) - f(x_t)\right) \leq \frac{1}{\gamma}\left(f(x_{t+1}) - f(x_0)\right) \leq \frac{1}{|\gamma|}f(x_0) - \frac{1}{|\gamma|}f(x_{t+1}) \leq \frac{1}{|\gamma|}f(x_o) < \infty$$

As advised in question:

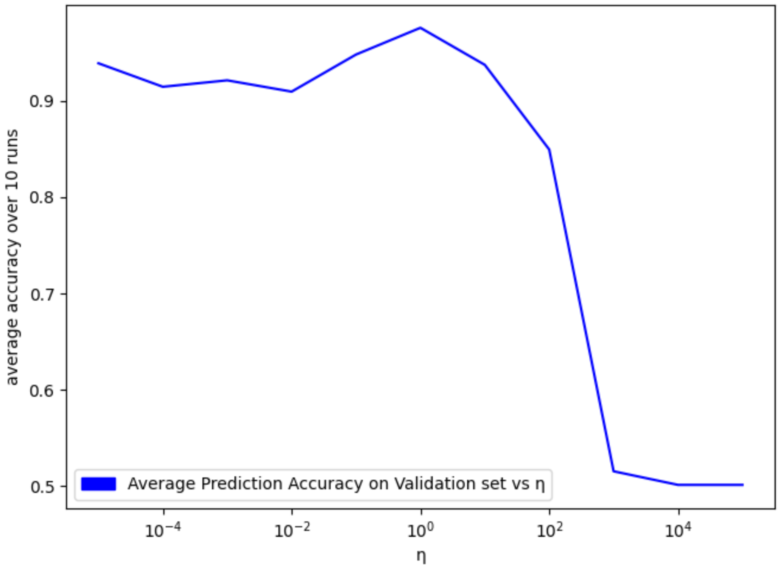
$$\sum_{t=0}^{\infty} a_t < 0 \to \lim_{t\to\infty} a_t = 0$$

$$\boldsymbol{\sum_{t=0}^{\infty}\|\nabla f(x_t)\|^2 < 0 \to \lim_{t\to\infty}\|\nabla f(x_t)\|^2 = 0}$$

Programming Assignment

1. Stochastic Gradient Decent (Hinge Loss)

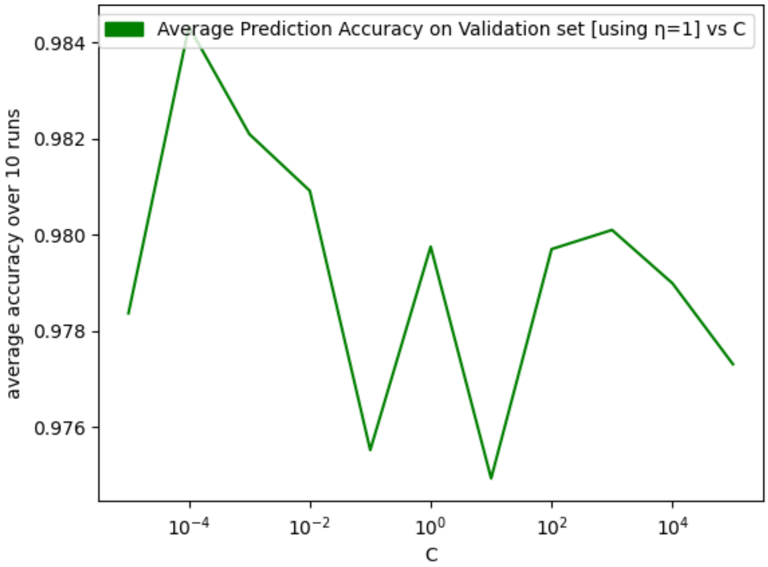
   a.



Finding best η from [1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000]

In Section A the best η was 1 with 97.7191% success on validation

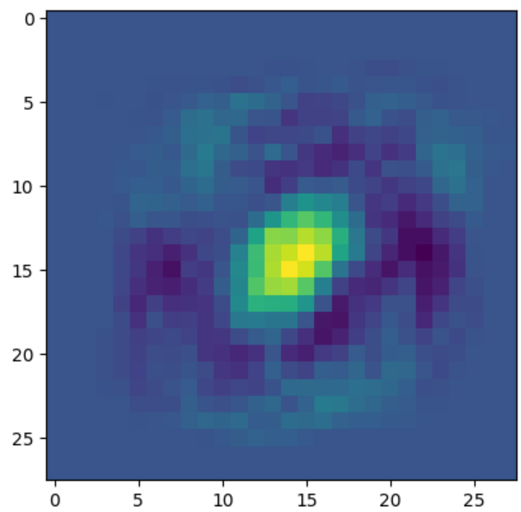
   b.



for η=1.00 Finding best C from [1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000]

In Section B the best C was 0.0001 with 98.4326% success on validation

   c. image representation of the classifier using best η and C:



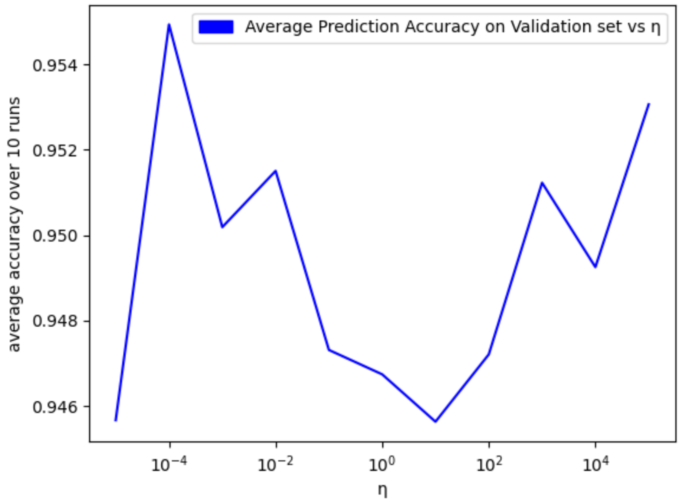
   d.

Section C&D: Finding Accuracy of η=1, C=0.0001
In Section C&D the best Accuracy was achieved with 99.1300% success on validation

2. Stochastic Gradient Decent (Log Loss)

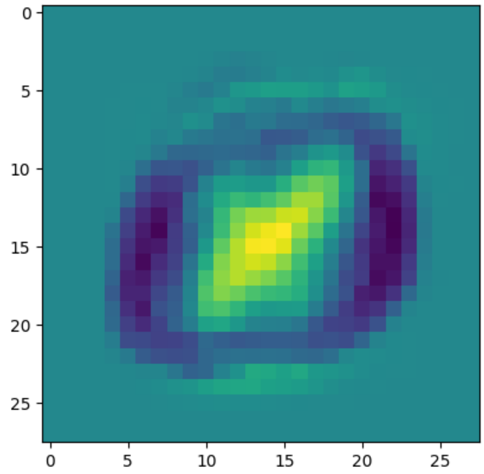
Section A: Finding best η from [1e-05, 0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000, 10000, 100000]

In Section A the best η was 0.0001 with 95.4936% success on validation



a.

In Section B the best w had 93.4493% success on validation



b.