Theoretical Assignment

1) **Step-size Perceptron.** Consider the modification of Perceptron algorithm with the following update rule:

$$w_{t+1} \leftarrow w_t + \eta_t y_t x_t$$

whenever $\hat{y}_t \neq y_t$ ($w_{t+1} \leftarrow w_t$ $otherwise$). Assume that data is separable with margin $\gamma > 0$ and that $\|x_t\| = 1$ 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 $\frac{4}{\gamma^2} log\left(\frac{1}{\gamma}\right)$. (Hint: use the fact that if $x \leq alog(x)$ then $x \leq 2alog(a)$). It's okay if you obtain a bound with slightly different constants, but the asymptotic dependence on γ should be tight.

Solution:

Assume w^* is a linear separator of the data and w.l.o.g $\|w^*\| = 1$.

(*) Moreover, assume $\gamma > 0$ is the margin of w^*, i.e., for every t it holds that $y_t \boldsymbol{w}^* \cdot \boldsymbol{x_t} = \frac{|w^* \cdot x_t|}{\|w^*\|} = dist\big(x_t, hyperplane(w^*)\big) \geq \gamma$

After each mistake at time t it holds that:

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

Therefore, after M mistakes (that made at first) we have:

$$(\boldsymbol{i}) \quad w_{t+1} \cdot w^* \geq w_M \cdot w^* + \frac{1}{\sqrt{M}}\gamma \geq \cdots \geq \gamma \sum_{i=1}^{M} \frac{1}{\sqrt{i}} \geq \gamma \int_1^{M+1} \frac{dx}{\sqrt{x}}$$
$$= \gamma\left(2\sqrt{M+1} - 2\right) \geq_{(**)} \gamma\sqrt{eM}$$

Where $(**)$ holds for M sufficiently large:

$$\lim_{M\to\infty} \frac{\sqrt{eM}}{2\sqrt{M+1} - 2} = \frac{\sqrt{e}}{2} < 1$$

In addition, after each mistake at time t it holds that:

$$\|w_{t+1}\|^2 = \|w_t + \eta_t y_t x_t\|^2 = \|w_t\|^2 + \underbrace{2\eta_t y_y w_t \cdot x_t}_{\substack{negative\ term\ since \\ we\ did\ a\ mistake}} + \|\eta_t y_t x_t\|^2$$
$$\leq \|w_t\|^2 + \eta_t^2 \|x_t\|^2 =_{\|x_t\|=1} \|w_t\|^2 + \frac{1}{t}$$

Therefore, after M mistakes (that made at first) we have:

$$(\boldsymbol{ii}) \quad \|w_{t+1}\|^2 \leq \|w_t\|^2 + \frac{1}{t} \leq \cdots \leq \sum_{i=1}^{M} \frac{1}{i} \leq 1 + \int_1^M \frac{dx}{x} = 1 + \ln(M) = \ln(eM)$$

By putting all together, we get:

$$\gamma\sqrt{eM} \leq_{(i)} w_{t+1} \cdot w^* \leq_{C-S} \|w_{t+1}\|\|w^*\| = \|w_t\| \leq_{(ii)} \sqrt{\ln(eM)}$$

Hence,

$$0 < \gamma\sqrt{eM} \leq \sqrt{\ln(eM)} \Rightarrow \gamma^2 eM \leq \ln(eM) \Rightarrow eM \leq \frac{1}{\gamma^2}\ln(eM)$$

By using the hint that mentioned in the question:

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

2) Convex functions.

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

Solution:

<u>Recall</u>: For convex set C we say that $f : C \to \mathbb{R}$ is convex if $\forall \alpha \in [0,1], \forall x_1, x_2 \in C$: $f(\alpha x_1 + (1-\alpha)x_2) \le \alpha f(x_1) + (1-\alpha) f(x_2)$.

First, notice that $\mathbb{R}^n$ is convex set. Let $x_1, x_2 \in \mathbb{R}^n$ and let $\alpha \in [0,1]$. It holds that:

$$g(\alpha x_1 + (1-\alpha)x_2) = f(A(\alpha x_1 + (1-\alpha)x_2) + b)$$
$$= f\big(\alpha(Ax_1 + b) + (1-\alpha)(Ax_2 + b)\big) \le_{convexity\ of\ f}$$
$$\le \alpha f(Ax_1 + b) + (1-\alpha) f(Ax_2 + b) = \alpha g(x_1) + (1-\alpha) g(x_2) \ \blacksquare$$

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

Solution:

Let $x_1, x_2 \in \mathbb{R}^n$ and let $\alpha \in [0,1]$. It holds that:

$$g(\alpha x_1 + (1-\alpha)x_2) = \max_i f_i(\alpha x_1 + (1-\alpha)x_2) \le_{covexity\ of\ each\ f_i}$$
$$\le \max_i \alpha f_i(x_1) + (1-\alpha) f_i(x_2) \le \max_i \alpha f_i(x_1) + \max_i (1-\alpha) f_i(x_2)$$
$$= \alpha g(x_1) + (1-\alpha) g(x_2) \ \blacksquare$$

(c) Let $\ell_{\log} : \mathbb{R} \to \mathbb{R}$ be the log loss, defined by $\ell_{\log}(z) = \log_2(1 + e^{-z})$. Show that $\ell_{\log}$ is convex, and conclude that the function $f : \mathbb{R}^d \to \mathbb{R}$ defined by $f(w) = \ell_{log}(y\boldsymbol{w} \cdot \boldsymbol{x})$ is convex with respect to $\boldsymbol{w}$.

Solution:

Recall (Calculus 1): Let $f: \mathbb{R} \to \mathbb{R}$ be a twice differentiable function. Then, f is convex if and only if $f''(x) \ge 0$.

$\ell_{\log}$ is differentiable twice and $\ell'_{\log}(z) = \ln(2)\frac{-e^{-z}}{1+e^{-z}}$

$$\Rightarrow \ell''_{\log} = \ln(2)\frac{e^{-z}(1+e^{-z}) - e^{-z}\cdot e^{-z}}{(1+e^{-z})^2} = \ln(2)\frac{e^{-z}(1+e^{-z}-e^{-z})}{(1+e^{-z})^2}$$
$$= \ln(2)\frac{e^{-z}}{(1+e^{-z})^2} \ge 0$$

Therefore, $\ell_{\log}$ is convex function on $\mathbb{R}$. In addition, for given $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$ we will show that $f_{x,y} : \mathbb{R}^d \to \mathbb{R}$ which is defined by $f(w) = \ell_{\log}(y\boldsymbol{w} \cdot \boldsymbol{x})$ is convex wrt w. Notice that $y\boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{w} \cdot (y\boldsymbol{x}) = A\boldsymbol{w}$ where $A = (y\boldsymbol{x})^t$. Therefore, by section (a) $\ell_{\log}(Aw) = f(w)$ is convex function. $\blacksquare$

3) Ranking.

(a) Consider the hinge loss function for the ranking objective problem:

$$\ell(h_{\boldsymbol{w}}(\overline{\boldsymbol{x}}), y) = \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - sgn(y_j - y_r)\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\}$$

We shell show that ℓ is convex wrt $\boldsymbol{w}$. Let $\boldsymbol{w_1}, \boldsymbol{w_2} \in \mathbb{R}^d$ and let $\alpha \in [0,1]$. It holds that,

$$\begin{aligned}
&\ell\left(h_{\alpha \boldsymbol{w_1} + (1-\alpha)\boldsymbol{w_2}}(\overline{\boldsymbol{x}}), y\right) \\
&= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - sgn(y_j - y_r)(\alpha \boldsymbol{w_1} + (1-\alpha)\boldsymbol{w_2}) \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\} \\
&= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\Big\{0, \alpha\left(1 - sgn(y_j - y_r)\boldsymbol{w_1} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right) \\
&\quad + (1-\alpha)\left(1 - sgn(y_j - y_r)\boldsymbol{w_2} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right)\Big\} \\
&\leq_{(*)} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\Big\{0, \alpha\left(1 - sgn(y_j - y_r)\boldsymbol{w_1} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right)\Big\} \\
&\quad + \max\Big\{0, (1-\alpha)\left(1 - sgn(y_j - y_r)\boldsymbol{w_2} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right)\Big\} \\
&= \alpha \ell\left(h_{\boldsymbol{w}_1}(\overline{\boldsymbol{x}}), y\right) + (1-\alpha)\ell\left(h_{\boldsymbol{w}_2}(\overline{\boldsymbol{x}}), y\right)
\end{aligned}$$

Where (*) is true α and $1-\alpha$ are non-negative numbers and we sum all potential non-negative terms in the max function. Therefore, ℓ is convex wrt $\boldsymbol{w}$. ■

(b) Consider the Kendall-Tau loss function for the ranking objective problem:

$$\Delta(y, y') = \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1}\{sgn(y'_j - y'_r) \neq sgn(y_j - y_r)\}$$

We shell show that the hinge loss upper-bounds the Kendall-Tau loss for linear classifiers: $h_w((x_1, \dots, x_k)) = (w \cdot x_1, \dots, w \cdot x_k)$.

Let $\boldsymbol{w} \in \mathbb{R}^d, \overline{\boldsymbol{x}} \in \mathcal{X}^k, \boldsymbol{y} \in \mathbb{R}^k$. It holds that,

$$\begin{aligned}
\Delta(h_{\boldsymbol{w}}(\overline{\boldsymbol{x}}), \boldsymbol{y}) &= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1}\{sgn(h_{\boldsymbol{w}}(\overline{\boldsymbol{x}})_j - h_{\boldsymbol{w}}(\overline{\boldsymbol{x}})_r) \neq sgn(y_j - y_r)\} \\
&= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1}\{sgn(\boldsymbol{w} \cdot \boldsymbol{x_j} - \boldsymbol{w} \cdot \boldsymbol{x_r}) \neq sgn(y_j - y_r)\} \\
&= \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1}\left\{sgn\left(\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right) \neq sgn(y_j - y_r)\right\} \\
&\leq_{(*)} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \mathbf{1}\left\{sgn\left(\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right) sgn(y_j - y_r) \in \{0, -1\}\right\} \\
&\leq_{(**)} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - sgn(y_j - y_r)\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\} = \ell(h_{\boldsymbol{w}}(\overline{\boldsymbol{x}}), y)
\end{aligned}$$

Where (*) is true since we took also cases which both sides are equal to zero. Part (**) is true since if $sgn\left(\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\right) sgn(y_j - y_r) = -1$ then $-sgn(y_j - y_r)\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r}) \geq 0$ and therefore $\max\{0, 1 - sgn(y_j - y_r)\boldsymbol{w} \cdot (\boldsymbol{x_j} - \boldsymbol{x_r})\} \geq 1$. In the other cases, we don't decrease the sum by taking 0 at the max function. ■

(c) Assume the data is linearly separable with margin $\gamma > 0$. We shell show now that minimizing the hinge loss will result a ranking function which minimizes the Kendall-Tau loss. Since the hinge loss upper-bounds the Kendall-Tau loss for linear classifiers, it suffices to show that the hinge loss can get arbitrarily close to Kendall-Tau loss. By the separability assumption, denote $w^* \in \mathbb{R}^d$ the true classifier such that $sgn(y_j^i - y_r^i)\boldsymbol{w}^* \cdot (\boldsymbol{x}_j^i - \boldsymbol{x}_r^i) \geq \gamma \; for \; all \; 1 \leq i \leq n \; and \; 1 \leq j < r \leq k$ (*).
Note that for every $c > 0$, $c\boldsymbol{w}^*$ yields the same classifier as $\boldsymbol{w}^*$. Therefore:

$$\sum_{\mathrm{i}=1}^{\mathrm{n}} \Delta\left(h_{c\boldsymbol{w}^*}(\overline{\boldsymbol{x}}^i), \boldsymbol{y}^i\right) \leq_{U-B} \sum_{\mathrm{i}=1}^{\mathrm{n}} \ell\left(h_{c\boldsymbol{w}^*}(\overline{\boldsymbol{x}}^i), \boldsymbol{y}^i\right)$$

$$= \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\left\{0, 1 - sgn(y_j^i - y_r^i)c\boldsymbol{w}^* \cdot (\boldsymbol{x}_j^i - \boldsymbol{x}_r^i)\right\}$$

$$\leq_{(*)} \sum_{\mathrm{i}=1}^{\mathrm{n}} \frac{2}{k(k-1)} \sum_{j=1}^{k-1} \sum_{r=j+1}^{k} \max\{0, 1 - c\gamma\} = n \cdot \max\{0, 1 - c\gamma\}$$

Taking $c \to \frac{1}{\gamma}$ and we get that $\sum_{\mathrm{i}=1}^{\mathrm{n}} \Delta\left(h_{c\boldsymbol{w}^*}(\overline{\boldsymbol{x}}^i), \boldsymbol{y}^i\right) \leq n \cdot \max\{0, 1 - c\gamma\} \to 0$

Hence, we minimize the K-T loss. ■

4) Gradient Decent on Smooth Functions.

f is β-smooth function, i.e., $\forall y, x \in \mathbb{R}^n$:

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

Substitute $y = x_{t+1}$ and $x = x_t$, and the iteration rule $x_t - x_{t+1} = \eta \nabla f(x_t)$. Then, for every $t \geq 0$:

$$0 \leq \eta\|\nabla f(x_t)\|^2 = \nabla f(x_t) \cdot \left(\eta \nabla f(x_t)\right) = \nabla f(x_t) \cdot (x_t - x_{t+1})$$

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

Consider the infinite sum of both non-negative sides, and we get:

$$\sum_{t=0}^{T} \|\nabla f(x_t)\|^2 \leq \frac{1}{\eta} \sum_{t=0}^{T} f(x_t) - f(x_{t+1}) + \frac{\beta}{2}\|x_t - x_{t+1}\|^2$$

$$= \frac{1}{\eta}\left(f(x_0) - f(x_{T+1})\right) + \frac{\beta}{2\eta} \sum_{t=0}^{\infty} \|\eta \nabla f(x_t)\|^2$$

So,

$$\left(1 - \frac{\beta\eta}{2}\right) \sum_{t=0}^{T} \|\nabla f(x_t)\|^2 \leq \frac{1}{\eta}\left(f(x_0) - f(x_{T+1})\right)$$

Since $\eta < \frac{2}{\beta}$, $1 - \frac{\beta\eta}{2} > 0$. Recall that f is non-negative function so $f(x_{T+1}) \geq 0$.

Therefore,

$$\sum_{t=0}^{T} \|\nabla f(x_t)\|^2 \leq \frac{1}{1 - \frac{\beta\eta}{2}} \frac{1}{\eta}\left(f(x_0) - f(x_{T+1})\right) \leq \frac{1}{1 - \frac{\beta\eta}{2}} \frac{1}{\eta} f(x_0)$$

Taking $T \to \infty$:

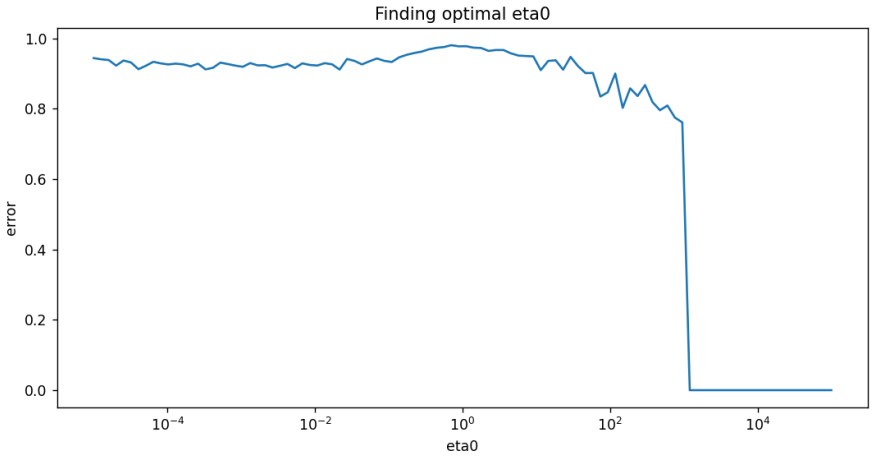
$$\sum_{t=0}^{\infty} \|\nabla f(x_t)\|^2 \leq \frac{1}{1 - \frac{\beta\eta}{2}} \frac{1}{\eta} f(x_0) < \infty$$

Therefore, $\lim_{t \to \infty} \|\nabla f(x_t)\|^2 = 0$ and also $\lim_{t \to \infty} \|\nabla f(x_t)\| = 0$. ■

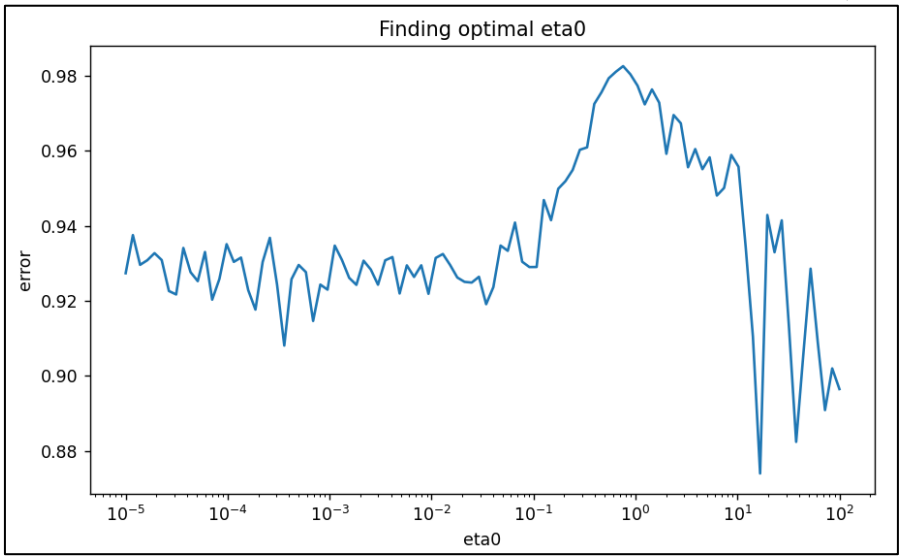
Programming Assignment

(1) SGD for Hinge loss

(a) The average accuracy on the validation set as a function of η_0:

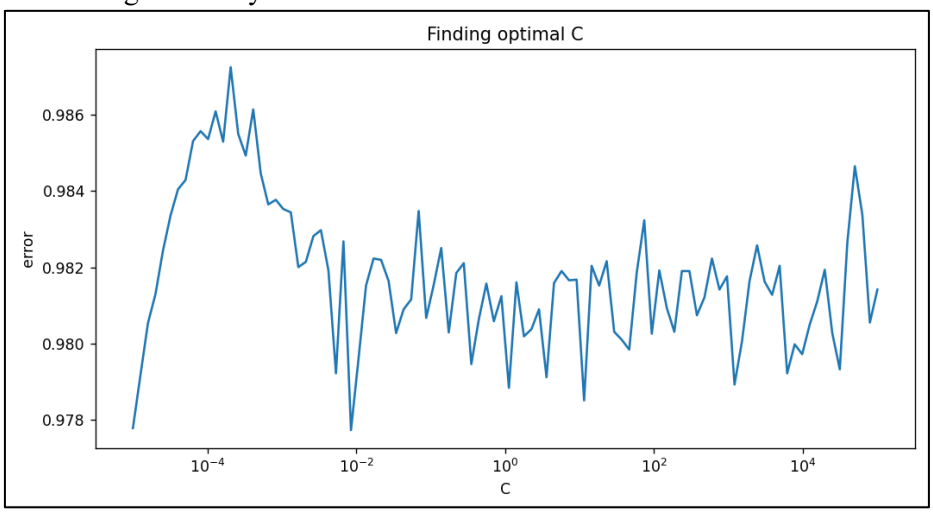


(First run)



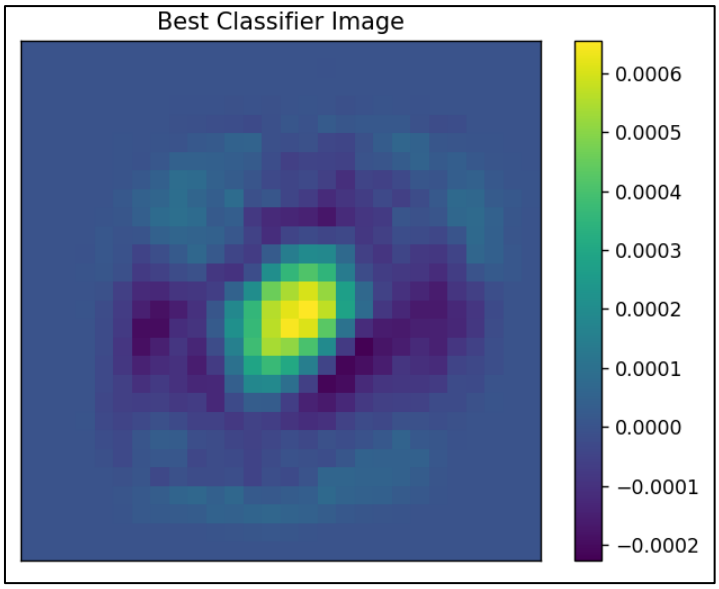
The best η_0: 0.7564

(b) The average accuracy on the validation set as a function of C:



The best C: 2.0565e-4

(c) An image of the trained classifier over the best η_0, C that were obtained earlier, with 20000 SGD iterations:



We can interpret the image by matching the bright areas (the green and yellow ones) with the digit 8 and the dark areas (the dark blue and black ones) with the digit 0. The classifier gives the most positives weights to digits with presentence in the center of the picture (in our case this is the middle of the digit 8); and gives the most negative weights to digits with presentence in the mid-sides of the image's center (in our case this is the sides of the digit 0). Moreover, we can see that the classifier almost ignores the positions of top and bottom parts of 8 and 0, as we can expect since it looks same at these areas.

(d) The accuracy of the best classifier on the test set: 0.9928.

(2) SGD for log-loss

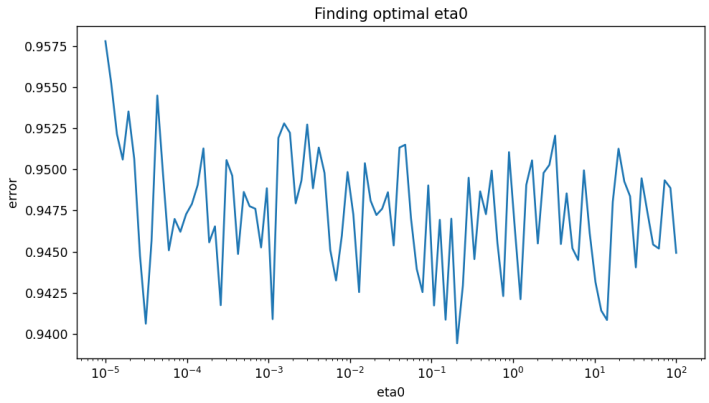
Consider the log-loss function: $\ell_{log}(\boldsymbol{w}, \boldsymbol{x}, y) = \log(1 + e^{-y\boldsymbol{w}\cdot\boldsymbol{x}})$. First, we shell derive the gradient updates of SGD for this objective function. The derivative of ℓ wrt $\boldsymbol{w}$ is:

$$\nabla\ell = \frac{-ye^{-y\boldsymbol{w}\cdot\boldsymbol{x}}}{1 + e^{-y\boldsymbol{w}\cdot\boldsymbol{x}}}\boldsymbol{x}$$

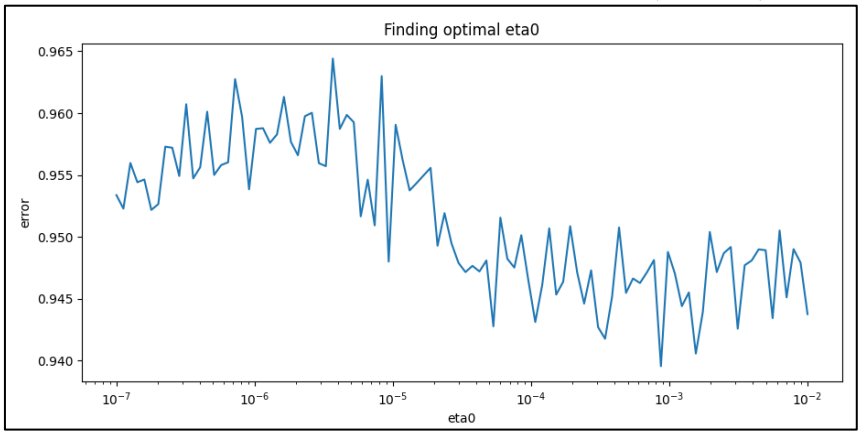
Start with $\boldsymbol{w_0}$. In each update step, we randomly choose $i \in \{1, \dots, n\}$ and denote the loss in x_i with classifier weight w as $f_i(w) = \log(1 + e^{-y_i\boldsymbol{w}\cdot\boldsymbol{x_i}})$. Then we get:

$$\boldsymbol{w_{t+1}} = \boldsymbol{w_t} - \eta_t\nabla f_i(\boldsymbol{w_t}) = \boldsymbol{w_t} + \eta_t\frac{y_i e^{-y_i\boldsymbol{w_t}\cdot\boldsymbol{x_i}}}{1 + e^{-y_i\boldsymbol{w_t}\cdot\boldsymbol{x_i}}}\boldsymbol{x_i}$$

(a) The average accuracy on the validation set as a function of η_0:

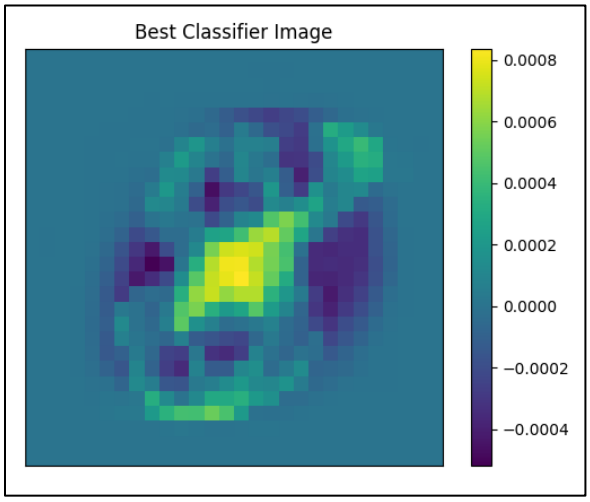


(First run)



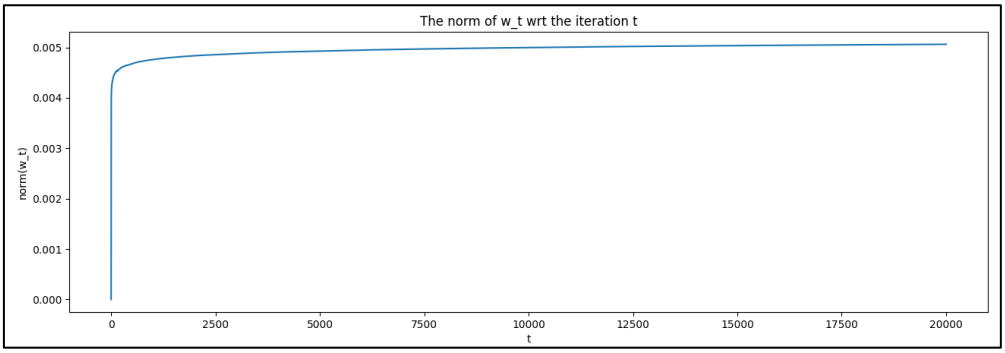
The best η_0: 3.274e-06

(b) An image of the trained classifier over the best that was obtained earlier, with 20000 SGD iterations:



The accuracy of the best classifier on the test set: 0.9851

(c) The norm of the classifier w_t as the iterations t growth:



As we can see, the norm of w_t get closer very fast to its optimum value (after ~130 iterations it reaches a norm of 0.0045, after 1000 iterations 0.00475 and after 20000 iterations it reaches a norm of 0.00505). We can conclude that as t is growth, w_t does not changes a lot. Recall that the SGD choose the parameter $\eta_t = \frac{\eta_0}{t}$ and the changes of the vector w is $\boldsymbol{w_{t+1}} = \boldsymbol{w_t} - \eta_t \nabla f_i(\boldsymbol{w_t})$. Therefore, it makes sense that $\boldsymbol{w_t}$ does not change a lot as t is big. The interesting phenomena is that indeed we obtain a very accurate result after few iterations.