Introduction to Machine Learning

Section 3

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 $\hat{y} \neq y$. Assume that data is separable with margin $\gamma > 0$ and that $||x_t|| = 1$ for all t. for any $1 \leq i \leq m$, Perceptron's i'th iterate takes the form:

$$w_{t+1}w^* = (w_t + \eta_t y_t x_t)w^* = w_t w^* + \underbrace{y_t w^* x_t}_{x_t y_t w^* x \geq \gamma} \frac{1}{\sqrt{t}} \geq w_t w^* + \frac{\gamma}{\sqrt{t}}$$

the M mistake hold: $w_M w^* \geq m\frac{\gamma}{\sqrt{m}} = \sqrt{m}\gamma$.
and now $||w_t||_2^2$ upper bounded is

$$M_\gamma \leq \frac{w^* \sum_{t=1}^m y_t x_t}{||w^*||} \leq \frac{w^* \sum_{t=1}^m (w_{t+1} - w_T)}{||w^*||\eta}$$

$$||w_{t+1}||_2^2 \leq \sqrt{\sum_{t=1}^m ||w_t + \eta_t y_t x_t||^2 - ||w_t||^2} \leq \sqrt{\sum_{t=1}^m \underbrace{2\eta_t y_t x_t}_{negtive} + \eta^2 ||x_t||^2}$$

$$\leq \sqrt{\sum_{t=1}^m \frac{1}{t}||x_t||} \leq \sqrt{H_m} \sim \log(\sqrt{m})$$

using Cauchy-Schwarz ineq

$$\gamma\sqrt{m} \leq w_M w^* \leq ||w_{t+1}||_2^2 \leq \log(\sqrt{m})$$

$$\Rightarrow \sqrt{m} \leq \frac{1}{\gamma}\log(\sqrt{m}) \Rightarrow \sqrt{m} \leq \frac{2}{\gamma}\log(\frac{1}{\gamma}) \Rightarrow m \leq \frac{4}{\gamma^2}\log^2(\frac{1}{\gamma})$$

2. Convex functions.

2.1

Let $f : R^n \to R$ a convex function, $A \in R^{n\times n}$ and $b \in R^n$ .for some $0 < \lambda < 1$,we like to have the graph of g on an interval $[x, y]$ falls below or on the graph. we can notice $b = \lambda b + (1 - \lambda)b$

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

using Jensen's inequality

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

and the sum of both convex function hold the convex property over R^n

2.2

Now lets consider $f_1(x), f_2(x) \dots f_m(x)$ convex function $f_i : R^d \to R$ and we will proof $g(x) = \max_i f_i(x)$ is also convex. using the property from section a $f_i(\lambda x + (1 - \lambda)y) \leq \lambda f_i(x) + (1 - \lambda)f_i(y)$, we take maximum of the both sides.

$$\max_i \{f_i(\lambda x + (1 - \lambda)y)\} \leq \max_i \{\lambda f_i(x) + (1 - \lambda)f_i(y)\}$$

$$\max_i \{f_i(\lambda x + (1 - \lambda)y)\} \leq \max_i \{\lambda f_i(x)\} + \max_i \{(1 - \lambda)f_i(y)\}$$

hence we can write $g(x)$ in the form

$$g(\lambda x + (1 - \lambda)y) \leq \lambda g(x) + (1 - \lambda)g(y).$$

$g(x)$ is convex

2.3

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

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

we know f is covex iff $f'' > 0$

$$\frac{d}{dz}\left(\log_2\left(1 + e^{-z}\right)\right) = -\frac{e^{-z}}{\ln(2)\left(1 + e^{-z}\right)}$$

$$\Rightarrow \frac{d}{dz}\left(-\frac{e^{-z}}{\ln(2)(1+e^{-z})}\right) = \frac{e^{-z}}{\ln(2)(1+e^{-z})^2} > 0$$

using section a,b. for $f(\mathbf{w})$ define by

$$f(\mathbf{w}) = \ell_{log}(y\mathbf{w}\cdot\mathbf{x}) = \sum_{i=1}^{n} \log_2(1+e^{-yx_iw})$$

lets set $f_i = \log_2(1+e^{-yx_iw})$. the set $\{f_i\}$ is convex set and any f_i can written as $f(\alpha x + (1-\alpha)y)$ hence $f(\mathbf{w})$ can written as

$$f(\mathbf{w}) = \sum_{i=1}^{n} f(\alpha x+(1-\alpha)y) \le n \max_i\{f_i\} \le n \max_i\{\lambda f_i(x)\}+n \max_i\{(1-\lambda)f_i(y)\}$$

3. GD with projection.

3.1

Let $y \in \mathbb{R}^d$ and $x = \prod_{\mathcal{K}}(y)$. and lets $z \in \mathcal{K}$ by assumption $\mathcal{K}$ is convex set, hence we can write any $k \in \mathcal{K}$
$(1-\lambda)x + \lambda z = x - \lambda(x-z) \in \mathcal{K}$ for any $\lambda \in (0,1)$

$$||x-y||^2 \le ||x-\lambda(x-z)-y||^2 = ||(x-y)-\lambda(x-z)||^2$$

$$\le ||x-y||^2 - 2\lambda\langle x-y, x-z\rangle + \lambda^2||x-z||^2$$

$$\Rightarrow \langle x-y, x-z\rangle \le \frac{\lambda}{2}||z-x||^2$$

the following hold for any $\lambda \in (0,1)$ since the right hand size can be small as we wish for a given z. on the other hand the right side can be less then 0 for some y s.t $y \notin \mathcal{K}$ and we get

$$\langle x-y, x-z\rangle \le 0$$

and now lets look at some $z \in \mathcal{K}$ and we choose some $\langle x-y, x-z\rangle \le 0$

$$||y-z||^2 - ||x-z||^2 = ||y-z+x-x||^2 - ||x-z||^2 =$$

$$||(x-z)-(x-y)||^2 - ||x-z||^2 = ||x-z||^2 - 2\langle x-y, x-z\rangle + ||x-y||^2 - ||x-z||^2 > 0$$

$$\Rightarrow ||y-z|| \ge ||x-z||$$

3.2

Theorem. *The GD with projection holds the Convergence Theorem. Given desired accuracy $\epsilon \geq 0$ set $\eta = \frac{B^2}{\epsilon}$ and ruining GD with projection for $T = \left(\frac{\epsilon G}{B}\right)^2$*

Proof. The GD with projection still holds the Convergence Theorem. using Jensen inequality and Convexity property

$$f(w) - f(w^*) \leq \frac{1}{T}\sum_{t=1}^{T} \nabla f(x_t)(x_t - x*) \leq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\eta}(x_t - y_{t+1})(x_t - x*) \leq$$

using the identity $2ab = ||a||^2 + ||b||^2 - ||a-b||^2$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2\eta}\left(||x_t - y_{t+1}||^2 + ||x_t - x^*||^2 - ||y_{t+1} - x^*||^2\right)$$

$$\leq \frac{1}{T}\sum_{t=1}^{T} \frac{1}{2\eta}\underbrace{\left(||x_t - x^*||^2 - ||y_{t+1} - x^*||^2\right)}_{3.1} + \frac{1}{T}\sum_{t=1}^{T} \frac{\nabla f(x_{t+1})}{2\eta}$$

using the result from 3.1 we know that $||y_{t+1} - x^*|| \geq ||x_{t+1} - x^*||$ and assuming $||\nabla f(x_{t+1})|| \leq G$ we get

$$\frac{||x_1 - x*||^2}{2\eta} + \frac{TG^2}{2\eta} \leq \frac{B^2}{2\eta} + \frac{TG^2}{2\eta}$$

for any $\epsilon \geq 0$ plug-in $\eta = \frac{B^2}{\epsilon}$ and $T = \left(\frac{\epsilon G}{B}\right)^2$

$$f(w) - f(w^*) \leq \frac{B^2}{2\eta} + \frac{TG^2}{2\eta} = \epsilon$$

□

4. Gradient Descent on Smooth Functions.

Let $f : \mathbb{R}^n \to \mathbb{R}$ be a β-smooth and non-negative function. we Consider the gradient descent algorithm applied on f with constant step size $\eta > 0$:

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

now lets compute x_t, x_{t+1} in f

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

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

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

since f is non-negative we can bound gradient squared norm of the gradient.

$$\sum_{t=1}^{k} ||\nabla f(x_t)||^2 \leq (\eta - \frac{\beta \eta^2}{2})^{-1} (f(x_1) - f(x_{k+1}))$$

Thus either the function values $f(x_k)$ tend to $-\infty$ or the sequence $\{||\nabla f(x_t)||^2\}$ is summable and therefore every limit point of the iterates x_k the GD is equal to zero, since $0 \leq f(x_{k+1}), f(x_k)$,and $\eta < 2/\beta$ lets define $f^* := \lim_{\to\infty} f(x_k)$

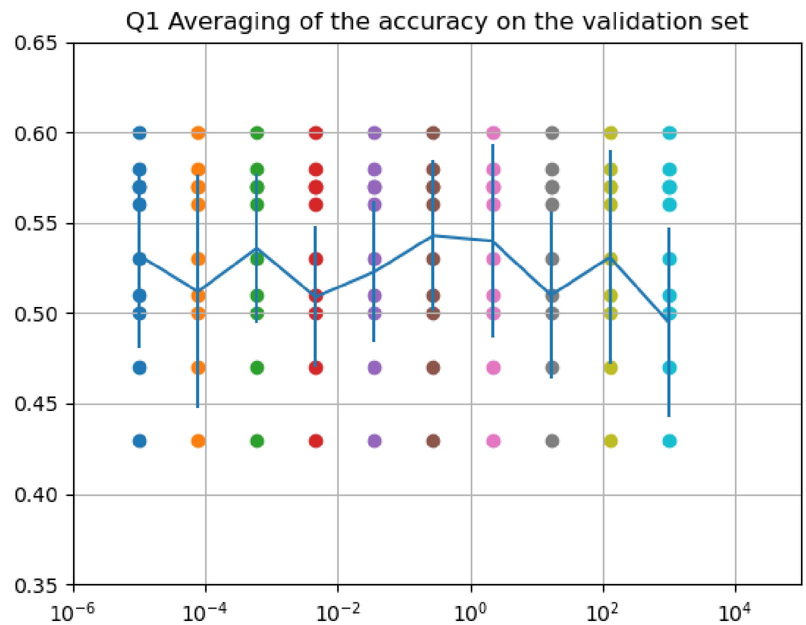
$$\min_t ||\nabla f(x_t)||^2 \leq \frac{1}{k} \sum_{t=1}^{k \to \infty} ||\nabla f(x_t)||^2 \leq \frac{1}{k} \frac{1}{\eta(1 - \frac{\beta \eta}{2})} (f(x_1) - f(x_{k+1})$$

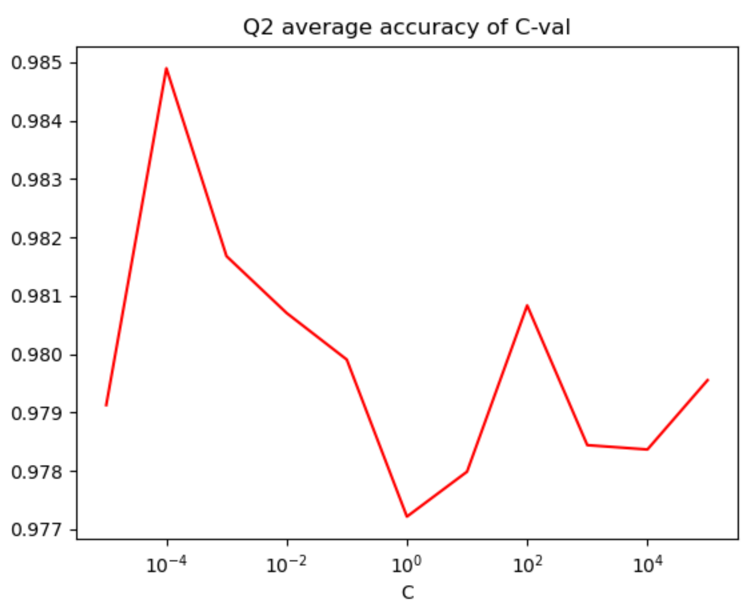
hence for some c (depends on x_1 value) we can set $c/\sqrt{k}$ that holds

$$||\nabla f(x_t)|| \leq \frac{c f(x_t + 1)}{\sqrt{t}} \xrightarrow[t \to \infty]{} 0$$

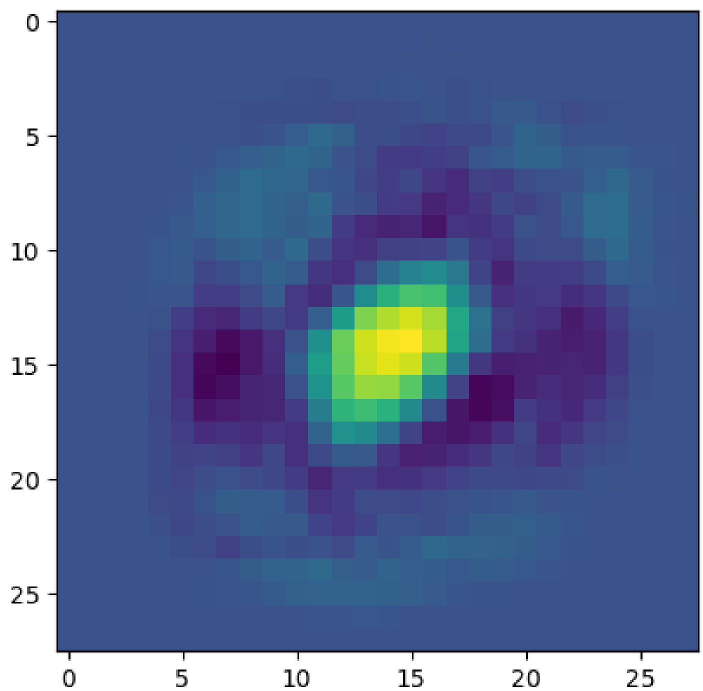
Programming Assignment

SGD for Hinge loss





(c)

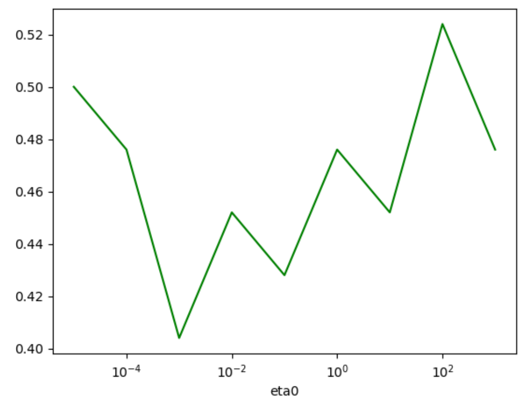


(d)

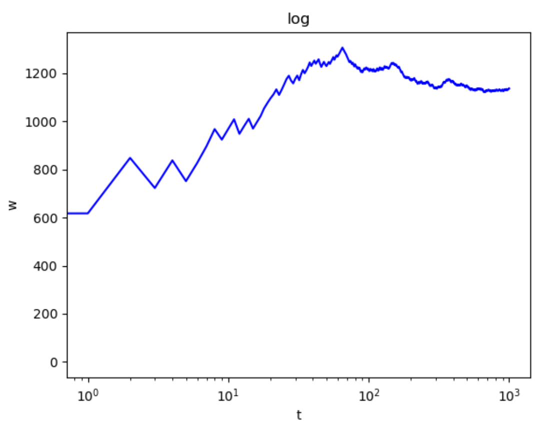
```
*****************************************************************
 the best classifier on the test set 0.9923234390992836
*****************************************************************
```

SGD for log-loss.

(a)



(b)



(c)

```
************************************************************************************
                      best accuracy  0.6033776867963152
************************************************************************************
```