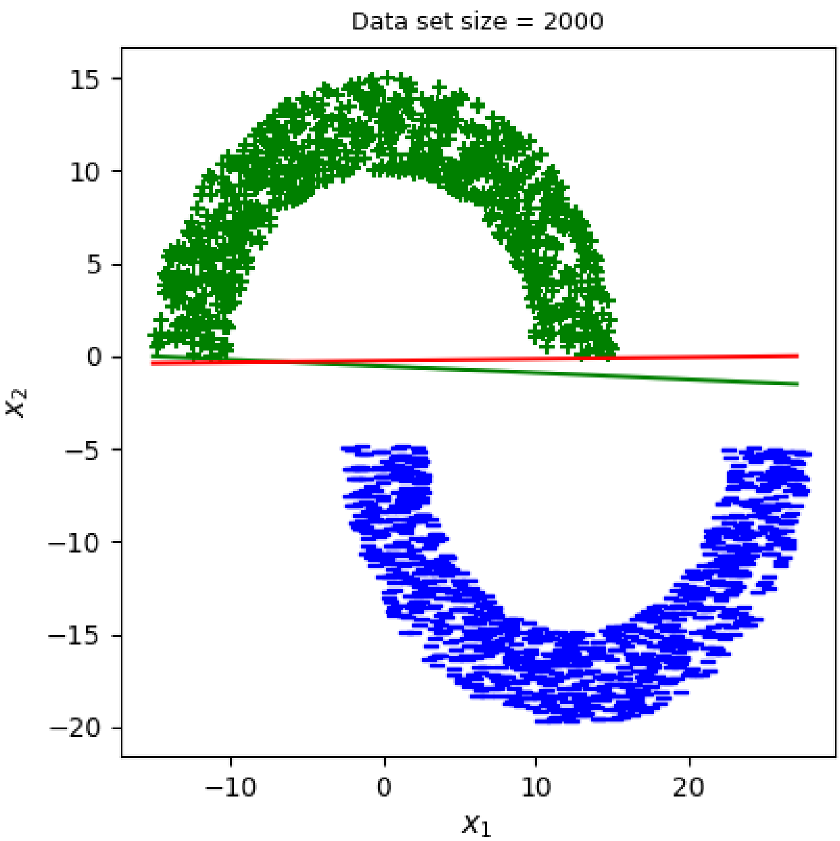
1.



Number of positive points: 958

Number of negatives points: 1042

Final correctness: 2000

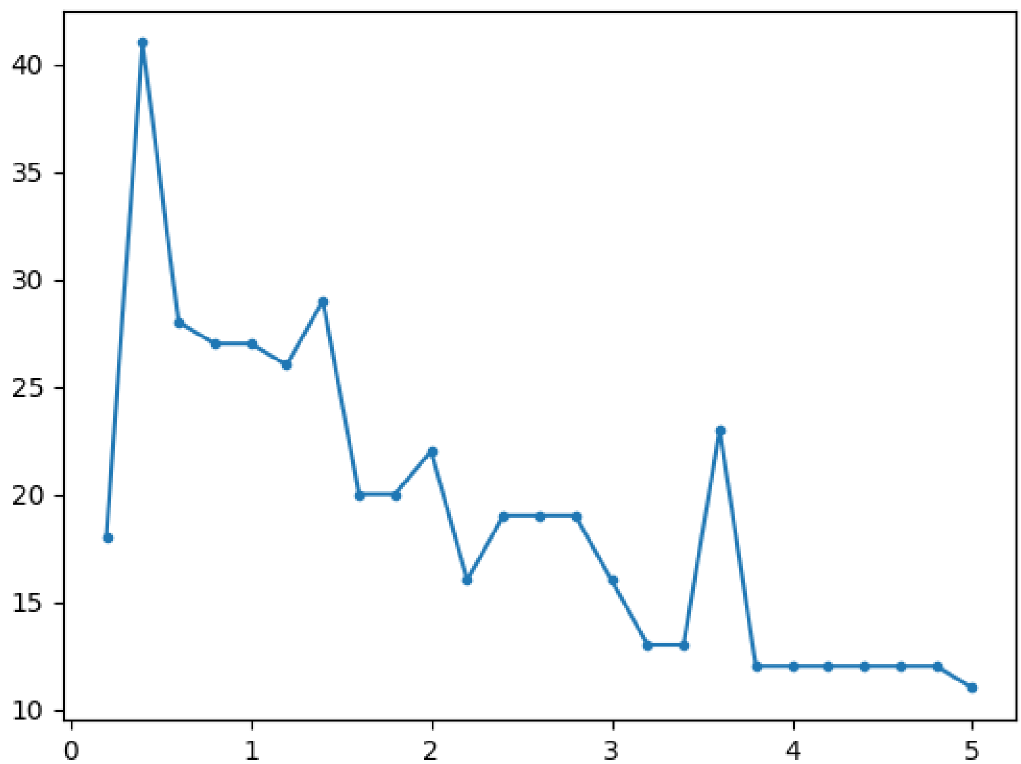
Total iteration: 31

Final w: [29. 1.92092337 53.66631267]

Liner regression coefficients: [0.24591466 -0.00937018 0.0790168]

From the graphs it can be inferred that PLA and linear regression achieves very close solution. Consider a hyperplane y = w0 + w1x1 + w2x2 for the -1 and 1 points given x in the space. Once we find the w through linear regression, we let w0 + w1x1 + w2x2 = 0, this hyperplane should separate the points approximately.

2.




```
Final correctness:  2000 . Total iteration:  18
Final w: [ 6.          -0.89097195 47.94025173]
Final correctness:  2000 . Total iteration:  41
Final w: [15.          -0.63429873 82.0719898 ]
Final correctness:  2000 . Total iteration:  28
Final w: [10.          -1.72886399 64.03266339]
Final correctness:  2000 . Total iteration:  27
Final w: [11.          -1.77381982 62.10530211]
Final correctness:  2000 . Total iteration:  27
Final w: [11.          -1.77381982 63.70530211]
Final correctness:  2000 . Total iteration:  26
Final w: [12.          -1.81877565 61.17794083]
Final correctness:  2000 . Total iteration:  29
Final w: [13.          -0.60062561 68.41241207]
Final correctness:  2000 . Total iteration:  20
Final w: [ 8.          -0.94113792 55.47314254]
Final correctness:  2000 . Total iteration:  20
Final w: [ 8.          -0.94113792 56.67314254]
Final correctness:  2000 . Total iteration:  22
Final w: [10.          -0.73920639 59.74861391]
Final correctness:  2000 . Total iteration:  16
Final w: [ 6.          -1.10571191 53.72224028]
Final correctness:  2000 . Total iteration:  19
Final w: [ 9.          -1.26842599 52.63027033]
Final correctness:  2000 . Total iteration:  19
```

```
Final w: [ 9.         -1.26842599 53.63027033]
Final correctness:  2000 . Total iteration:  19
Final w: [ 9.         -1.26842599 54.63027033]
Final correctness:  2000 . Total iteration:  16
Final w: [ 6.          0.13843134 55.84381562]
Final correctness:  2000 . Total iteration:  13
Final w: [ 5.         -0.39078826 46.52547204]
Final correctness:  2000 . Total iteration:  13
Final w: [ 5.         -0.39078826 47.32547204]
Final correctness:  2000 . Total iteration:  23
Final w: [13.          0.87921332 61.61639105]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 43.89948258]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 44.49948258]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 45.09948258]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 45.69948258]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 46.29948258]
Final correctness:  2000 . Total iteration:  12
Final w: [ 6.         -0.45419319 46.89948258]
Final correctness:  2000 . Total iteration:  11
Final w: [ 7.         -0.51759811 41.27349311]
```

It generally takes more iterations for PLA to converge when *sep* is small. Looks like ||w|| is increasing when *sep* is decreasing. The R term is fixed due to x are fixed. The p is the minimum of $y_n w^t x_n$, which is less affected by the change of ||w||. So, the overall effect is to increase the time to converge for PLA.

3.

3) a) Using taylor's expansion of e^x,

where $e^x = 1 + x + x^2/2! + x^3/3! \cdots$

we have,

$$E(u+\Delta u, v+\Delta v) = e^{u+\Delta v} + e^{2(v+\Delta v)} + e^{(u+\Delta u)(v+\Delta v)}$$

$$+ (u+\Delta u)^2 - 3(u+\Delta u)(v+\Delta v) + 4(v+\Delta v)^2 - 3(u+\Delta u) - 5(v+\Delta v)$$

$$= \left(1 + (u+\Delta u) + \frac{(u+\Delta u)^2}{2}\right) + \left(1 + 2(v+\Delta v) + \frac{4(v+\Delta v)^2}{2}\right) + \left(1 + (u+\Delta u)(v+\Delta v) + \frac{(u+\Delta u)^2 (v+\Delta v)^2}{2}\right)$$

$$+ (u+\Delta u)^2 - 3(u+\Delta u)(v+\Delta v)$$

$$+ 4(v+\Delta v)^2 - 3(u+\Delta u) - 5(v+\Delta v)$$

At $(u,v) = (0,0)$, take the first order taylor's expansion of E, we have

$$E(u,v) \approx \hat{E}_1$$

$$= 1 + \Delta u + 1 + 2\Delta u + 1 - 3\Delta u - 5\Delta v$$

$$= -2\Delta u - 3\Delta v + 3$$

M T W T F S S
Page No.:
Date:
YOUVA

So, $a_u = -2$, $a_v = -3$, and $a = 3$

b) To satisfy $||(\Delta u, \Delta v)|| = 0.5$

let $\Delta u = 0.5 \cos\theta$ and
$\Delta v = 0.5 \sin\theta$.

then $E(u,v) = -\cos\theta - 1.5 \sin\theta + 3$

Take derivative w.r.t θ and set it to 0

we have, $\sin\theta - 1.5\cos\theta = 0$
$\sin\theta = 1.5 \cos\theta$

Take $\sin\theta^2 + \cos\theta^2 = 1$, we have $\cos\theta = \frac{2}{\sqrt{13}}$

and $\sin\theta = \frac{3}{\sqrt{13}}$

The optimal $(\Delta u, \Delta v)$ is thus $\left(\frac{1}{\sqrt{13}}, \frac{3}{2\sqrt{13}}\right)$

The resulting $E(u + \Delta u, v + \Delta v)$ is computed below using the optimal $(\Delta u, \Delta v)$ here:

Now we compute the gradient $\nabla E(u,v)$

$$\nabla E(u,v) = \left[\frac{\delta E}{\delta u}, \frac{\delta E}{\delta v}\right]^T$$

$$= \begin{bmatrix} e^u + ve^{uv} + 2u - 3v - 3 \\ 2e^{2v} + ue^{uv} - 3u + 8v - 5 \end{bmatrix}$$

At point $(u,v) = (0,0)$, we have $\nabla E(0,0) = (-2, -3)$, The optimal $(\Delta u, \Delta v)$ is parallel to the negative gradient direction, i.e. $-\nabla E(0,0) = (2,3)$ which is consistent with the results computed above.

c) Approx. $E(u+\Delta u, v+\Delta v)$ by $\hat{E}_2(\Delta u, \Delta v)$ around $(u, v) = (0,0)$ we have

$$\hat{E}_2(\Delta u, \Delta v) = \left(1 + \Delta u + \frac{1}{2}\Delta u^2\right)$$

$$+ \left(1 + 2\Delta v + 2\Delta v^2\right) +$$

$$(1 + \Delta u \Delta v) +$$

$$\Delta u^2 - 3\Delta u \Delta v + 4\Delta v^2$$

$$-3\Delta u - 5\Delta u$$

$$= \frac{3}{2}\Delta u^2 + 6\Delta v^2 - 2\Delta u \Delta v -$$

$$2\Delta u - 3\Delta v + 3$$

So, $b_{uu} = \frac{3}{2}$, $b_{vv} = 6$, $b_{uv} = -2$

$b_u = -2$, $b_v = -3$, $b = 3$

d) Take derivatives of $\hat{E}_2(\Delta u, \Delta v)$ w.r.t. Δu and let them equal to 0, we have

$$\frac{\delta \hat{E}_2}{\delta \Delta u} = 3\Delta u - 2\Delta v - 3 = 0$$

$$\frac{\delta \hat{E}_2}{\delta \Delta v} = 12\Delta v - 2\Delta u - 3 = 0$$

M T W T F S S
Page No.:
Date:
YOUVA

writing the ~~above~~ equations in matrix form we have

$$\begin{bmatrix} \frac{\delta \hat{E}_2}{\delta \Delta u} \\ \frac{\delta \hat{E}_2}{\delta \Delta v} \end{bmatrix} = \begin{bmatrix} 3 & -2 \\ -2 & 12 \end{bmatrix} \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$$

$$= -\nabla E(0,0)$$

Solving the functions we get

$$(\Delta u^*, \Delta v^*) = \left(\frac{30}{32}, \frac{13}{32}\right)$$

We compute $\nabla^2 E(u,v)$.

Take the $\nabla E(u,v)$ computed above in problem (b) we have,

$$\nabla^2 E(u,v) = \begin{bmatrix} \frac{\delta \nabla E(u,v)}{\delta u} \\ \frac{\delta \nabla E(u,v)}{\delta v} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\delta \nabla E_u(u,v)}{\delta u} & \frac{\delta \nabla E_u(u,v)}{\delta v} \\ \frac{\delta \nabla E_v(u,v)}{\delta u} & \frac{\delta \nabla E_v(u,v)}{\delta v} \end{bmatrix}$$

$$= \begin{bmatrix} e^u + v^2 e^{uv} + 2 & e^{uv} + uve^{uv} - 3 \\ e^{uv} + uve^{uv} - 3 & 4e^{2v} + u^2 e^{uv} + 8 \end{bmatrix}$$

At point $(u,v) = (0,0)$ we have

$$\nabla^2 E(0,0) = \begin{bmatrix} 3 & -2 \\ -2 & 12 \end{bmatrix}$$

we can see that this $\nabla^2 E(0,0)$ is the same matrix as above when we take derivative of $\hat{E}_2(\Delta u, \Delta v)$ w.r.t. Δu and Δv

The eq. becomes

$$\begin{bmatrix} \frac{\delta \hat{E}_2}{\delta \Delta u} \\ \frac{\delta \hat{E}_2}{\delta \Delta v} \end{bmatrix} = \nabla^2 E(0,0) \begin{bmatrix} \Delta u \\ \Delta v \end{bmatrix}$$

$$= -\nabla E(0,0)$$

Since $\nabla^2 E(0,0)$ is positive definite we can solve for the optimal $(\Delta u, \Delta v)$ and have

$$\begin{bmatrix} \Delta u^* \\ \Delta v^* \end{bmatrix} = -\left(\nabla^2 E(0,0)\right)^{-1} \nabla E(0,0)$$

solve the above equation, we obtain the same solution as

$$(\Delta u^*, \Delta v^*) = \left(\frac{30}{32}, \frac{13}{32}\right)$$

e) From the numerical program, the calculatio
$\hat{E}_1$ is not approximation to $E(u+\Delta u, v-$
since it has larger difference compa
to $\hat{E}_2$ approx.

Newton direction is very close to the optimal direction obtained by minimizi
$E(u+\Delta u, v+\Delta v)$. The value compu
with Newton direction is just a little bit larger than the minimal value.

e.) Calculations using the program:

```
E(u+du,v+dv) with (du,dv) from E_1 approx. = 2.2508597349929693

Newton direction: (0.4587778126549621, 0.19880371881715023)

E(u+du,v+dv) = 1.8904907903020918

Optimal direction by minimizing E(u+du,v+dv): (0.4355689881974021,
0.2455191571358361)

Minimal E(u+du, v+dv): 1.8684370301391746
```

4.

5.

Consider a given H

If the best approximation from H is less complex than the initial target function, then when we increase the complexity of f, the deterministic noise in general should increase, since it'll be harder for functions in H to fit the target function. There'll be a higher tendency to overfit.

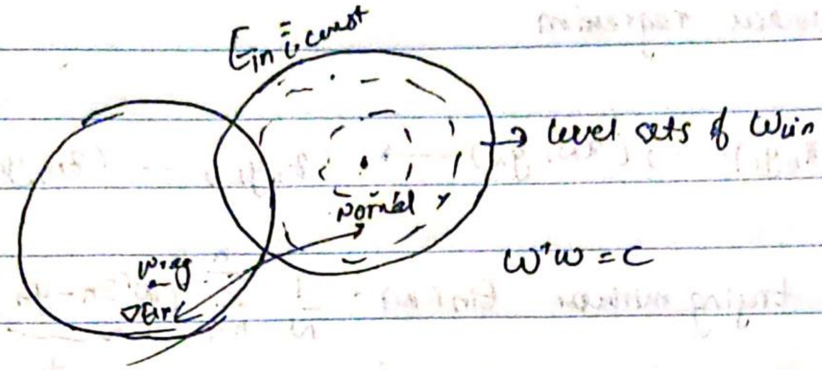
If the best approximation from H is more complex than the initial target function, then when we increase the complexity of f, the deterministic noise in general may decrease first, reducing the deterministic noise and there'll be a lower tendency to overfit. But once the complexity of f exceeds the best function approximation from H, and if we continue increase the complexity of f, we will increase the deterministic noise and thus increase the tendency to overfit.

(b) Given a fixed f

If the best approximation from H is less complex than the target function, then when we decrease the complexity of H, we increase the deterministic noise thus increasing the tendency of overfit.

If the best approximation from H is more complex than the target function, then when we decrease the complexity of H, we will decrease the deterministic noise thus decreasing the tendency of overfit. Well, if we continue to decrease the complexity of H, passing the point where its complexity is equal to f, we start to increase the deterministic noise again and thus increasing overfit.

6.



→ The surface $w^T w = c$ at optimal we should be $\perp^{er}$ to ∇E_{in},

when surface is $\perp^{er}$ the E_{in}, it is also $\parallel^{el}$ to normal

$\Rightarrow \nabla E_{in} \propto -w_{reg}$

If w_{reg} need to be optimal, then for some parameter λ_C

$$\nabla E_{in}(w_{reg}) = -2\lambda_C w_{reg}$$

$$\nabla E_{in}(w_{reg}) + 2\lambda_C w_{reg}$$

Taking gradient common

$$\nabla E_{in}(w) + \lambda_C \frac{w^T w}{2N} \Big|_{w = w_{reg}} = 0$$

linear regression

Aug 6)

$(x_1, y_1) \cdots, (x_N, y_N) \longrightarrow (z_1, y_1, \cdots (z_N, y_N)$

trying minimise $E_{in}(w) = \frac{1}{N} \sum_{n=1}^{n} \underbrace{(w^T z_n - y_n)}_{\text{minimise this}}^2$

When obtained, w_{lin} is free from constrained, the state is called unregularized.

for regularization $\underset{\text{wts.}}{w^T w} \le \underset{\text{constant}}{C}$

This is referred as soft order constraint as it pushes the wts. to small in size with changing order

$w_{reg} \in H_C$ ← regularized wts.

w_{reg} is calculated using such regularization parameters in a way that in sample error is minimised

In sample error for a transformed eqn

$$E_{in}(w) = \frac{(Zw - y)^T (Zw - y)}{N}$$

Hence $(Zw - y)^T$ & $(Zw - y)$ are cost curves that define radius of w_{lin}.

Now picking a C & minimizing $E_{in}(w)$ subject to:

$$w^T w \le C$$

$$E_{aug}(w) = E_{in}(w) + \lambda_c w^T w$$

because C increases, λ_c decreases.

$$E_{aug}(w) = \frac{1}{N}\left[(Zw - y)^T (Zw - y) + \lambda_c w^T w) \right]$$

$$\lambda = \frac{\lambda_c}{N}$$

$$E_{aug} = \frac{(Zw - y)^T (Zw - y) + \lambda w^T w}{N}$$

obtained by pretrial

~~QA~~ $\nabla E_{aug}(w) = 2Z^T(Zw - y) + 2\lambda w = 0$

$$= 2(Z^T Z + \lambda I) w - 2 Z^T y = 0$$

$$(Z^T Z + \lambda I) w = Z^T y$$

$$w_{reg} = (Z^T Z + \lambda I)^{-1} Z^T y.$$