Prob 1)

1.1 1) $P(x) = \dfrac{e^{\frac{-x^2}{2c^2}}}{\sqrt{2\pi c^2}}$

$$\Rightarrow H(x) = -\int_{-\infty}^{\infty} P(x) \log P(x)\, dx$$

$$= -\int_{-\infty}^{\infty} \left[\frac{e^{\frac{-x^2}{2c^2}}}{\sqrt{2\pi c^2}} \log_e \frac{e^{-x^2/2c^2}}{\sqrt{2\pi c^2}} \right] dx$$

$$= -\int \frac{e^{-\frac{x^2}{2c^2}}}{\sqrt{2\pi c^2}} \left[\frac{-x^2}{2c^2} - \log_e \sqrt{2\pi c^2} \right] dx$$

$$= \underbrace{\frac{1}{2c^2\sqrt{2\pi c^2}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2c^2}\, dx}_{b} - \underbrace{\frac{\log_e \sqrt{2\pi c^2}}{\sqrt{2\pi c^2}} \int_{-\infty}^{\infty} e^{-x^2/2c^2}\, dx}_{=a}$$

For a

We know, $\int_{-\infty}^{\infty} e^{-x^2/2c^2}\, dx = \sqrt{2\pi c^2}$

$$a = -\frac{\log_e \sqrt{2\pi c^2}}{\sqrt{2\pi c^2}} \int_{-\infty}^{\infty} e^{-x^2/2c^2}\, dx = -\frac{\sqrt{2\pi c^2}}{\sqrt{2\pi c^2}} \log_e \sqrt{2\pi c^2}$$

$$= -\log_e \sqrt{2\pi c^2}$$

$$(b) = \frac{1}{2c^2\sqrt{2\pi c^2}} \int_{-\infty}^{\infty} x^2 e^{-x^2/2c^2} dx$$

Now using, $\int_{-\infty}^{\infty} x^2 e^{-x^2} dx = \frac{\sqrt{\pi}}{2}$

$$(b) = -\frac{1}{2}$$

So, $H(x) = -(a+b) = \frac{1}{2} + \log\sqrt{2\pi c^2}$

104

PDF

4) $f_X(x) = P(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$

variation for also PDF

$y = f(x) = \delta(x-\theta) = \begin{cases} +\infty, & x = \theta \\ 0, & x \neq \theta \end{cases}$

$$\int_{-\infty}^{\infty} \delta(x-\theta)\, dx = 1$$

$y = \delta(x-\theta) = \begin{cases} a \text{ ; if } & x = \theta \\ 0 \text{ ; if } & x \neq \theta \end{cases}$

$f_Y(y) = P(Y=y) = \begin{cases} 1 & \text{if } y = a \\ 0 & \text{if } y = 0 \end{cases}$

So we are almost sure of $P(Y)$ distribution.

Therefore $H(f_Y(y)) = -\infty$ (minimum

2.1

$$f(x) = \frac{2a}{1+e^{-bx}} - a$$

(1)

$$f'(x) = \frac{2ab\, e^{-bx}}{(1+e^{-bx})^2}$$

$$= \left(\frac{2a}{1+e^{-bx}}\right)\left(\frac{b\, e^{-bx}}{1+e^{-bx}}\right)$$

$$= \underbrace{\left(\frac{2a}{1+e^{-bx}}\right)}_{\parallel} \underbrace{\frac{b}{2a}}_{\parallel} \underbrace{\left(2a - \frac{2a}{1+e^{-bx}}\right)}_{\parallel}$$

$$= (a + f(x)) \frac{b}{2a} (a - f(x))$$

$$f'(x) = \frac{b}{2a}\left(a^2 - f(x)^2\right)$$

2.2

(2)

$$f''(x) = \frac{d f'(x)}{dx} = \frac{b}{2a}\left(-2 f(x) f'(x)\right)$$

$$= -\frac{b}{a} f(x) f'(x)$$

$$\Rightarrow f(x) = \frac{2a}{1+e^{-bx}} - a \; ; \quad f(-\infty) = \frac{2a}{1+e^{bx}} - a = \underline{\underline{-a}}$$

$$f(0) = \frac{2a}{1+e^{-b(0)}} - a = \underline{\underline{0}} \; ; \quad f(\infty) = \frac{2a}{1+e^{-b\infty}} - a = \underline{\underline{a}}$$

$$\Rightarrow f'(x) = \frac{b}{2a}\left(a^2 - f(x)^2\right)$$

$$f'(-\infty) = \frac{b}{2a}\left(a^2 - f^2(-\infty)\right) = \frac{b}{2a}\left(a^2 - a^2\right) = \cancel{\frac{b}{2}(a+1)} \; 0$$

$$f'(0) = \frac{b}{2a}\left(a^2 - 0\right) = \frac{ab}{2} \; ; \; f'(\infty) = \frac{b}{2a}\left(a^2 - a^2\right) = \cancel{\frac{b}{2}(a-1)} \; 0$$

$$\Rightarrow f''(x) = -\frac{b}{a} f(x) f'(x)$$

$$f''(-\infty) = \frac{-b}{a} f(-\infty) f'(-\infty)$$

$$\frac{-b}{a} \cdot \left[\frac{b}{2}(a+1)\right] \cdot (-a)$$

$$f''(0) = -\frac{b}{a} f(0) f'(0)$$

$$= 0$$

$$f''(\infty) = \frac{-b}{a} f(\infty) f(\infty)$$

$$= \frac{-b}{a} \cdot a \cdot \frac{b}{2}(a-1)$$

2.3

$$\Rightarrow f''(x) = \frac{-b}{a} f(x) f'(x)$$

$$f'''(x) = \frac{-b}{a} f'(x)^2 - \frac{b}{a} f(x) f''(x) = 0$$

$$= -\frac{b}{a} f'(x)^2 \; \frac{-b}{a} f(x) \left[\frac{-b}{a} f(x) f'(x)\right]$$

$$= -\frac{b}{a} f'(x)^2 + \frac{b}{a} f'(x) \left(\frac{b}{a} f(x)^2\right)$$

And also,

$$f'(x) = \frac{b}{2a}\left(a^2 - f(x)^2\right) = \frac{ab}{2} - \frac{1}{2}\frac{b}{a} f(x)^2$$

$$\frac{b}{a} f(x)^2 = 2 \cdot \frac{ab}{2} - 2 \cdot f'(x) = ab - 2 f'(x)$$

$\Rightarrow$ Then $f'''(x) = \frac{-b}{a} f'(x)^2 + \frac{b}{a} f'(x)\left(ab - 2 f'(x)\right)$

$$= \frac{-b}{a} f'(x)^2 + b^2 f'(x) - 2\frac{b}{a} f'(x)^2$$

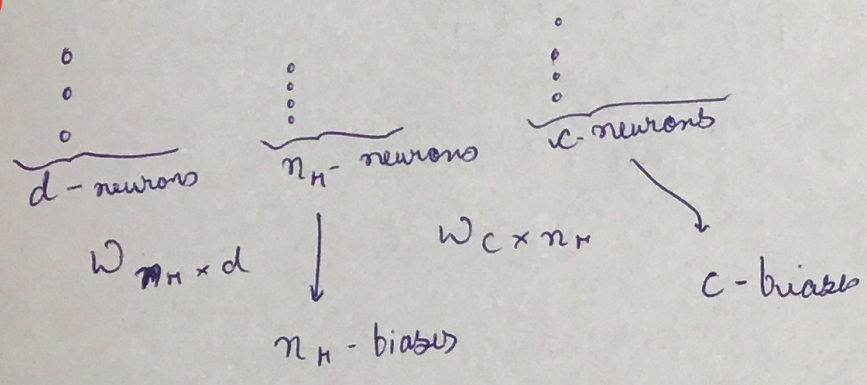
$$= -\frac{3b}{a} f'(x)^2 + b^2 f'(x)$$

$$f'''(x) = \left(\frac{-3b}{a} f'(x) + b^2\right) f'(x) = 0$$

Thus, $x = -\infty, \infty$ as, $f'(x) = 0$, at $x = -\infty, \infty$

Prob. 3.1

3)



Total learnable Parameters $= n_H \times d + n_H + c \times n_H + c$

3.2

2) $f(w_2 f_2(w_1 x)) = f(-w_2 f(-w_1 x))$

To proof

First lets see, $f(-x) = -f(x)$

$$f(x) = \frac{a(1 - e^{bx})}{(1 + e^{bx})}$$

$$f(-x) = \frac{a\left(1 - \frac{1}{e^{bx}}\right)}{\left(1 + \frac{1}{e^{bx}}\right)} = -\frac{a(1 - e^{bx})}{(1 + e^{bx})}$$

$$= -f(x)$$

Thus, $f(-w_2 f(-w_1 x))$

$$= f(-w_2 \times -f(w_1 x))$$

$$= f(w_2 f(w_1 x))$$

Prob 4

4.01

(1) We know,

$$\gamma_k = \frac{v(x^k) - v(x^{k+1})}{v(x^k)} = \alpha_k \left(\underbrace{\frac{g^{k^T} Q g^k}{g^{k^T} Q^{-1} g^k}}_{\max} \right) \left(\frac{2 g^{k^T} g^k}{g^{k^T} Q g^k} - \alpha_k \right)$$

min

where, $g^k = Q(x^k - x^*)$.

$\Rightarrow$ Lower bound on γ_k using property of SPD matrix:-

$$\gamma_k \geq \frac{\lambda_{min}(Q)}{\lambda_{max}(Q^{-1})} \left(\frac{2}{\lambda_{max}(Q)} - \alpha_k \right)$$

$$\gamma_k \geq \lambda^2_{min}(Q) \left(\frac{2}{\lambda_{max} Q} - \alpha_k \right).$$

We want $\gamma_k > 0$ such that:- $\sum_{k=0}^{\infty} \gamma_k = \infty$

Therefore $\frac{2}{\lambda_{max}(Q)} - \alpha_k > 0$

$$\alpha_k < \frac{2}{\lambda_{max}(Q)}$$

4.2

2 ⇒ For learning time calculation, its better to look for order of convergence.

Now,

$$\gamma^k = \frac{V(x^k) - V(x^{k+1})}{V(x^k)}$$

$$\Rightarrow \underbrace{1 - \gamma^k}_{\text{Want upper bound on this}} = \frac{V(x^{k+1})}{V(x^k)} \quad \begin{matrix} \leftarrow \text{upperbound} \\ \leftarrow \text{Lower bound} \end{matrix}$$

⇒ Now,

$$V(x^{k+1}) = \frac{1}{2}(x^{k+1} - x^*)^T Q (x^{k+1} - x^*)$$

$$\leq \frac{1}{2} \lambda_{max}(Q) \| x^{k+1} - x^* \|_2^2$$

Also

$$V(x^{k+1}) \geq \frac{1}{2} \lambda_{min}(Q) \| x^{k+1} - x^* \|^2$$

⇒ Thus, from upper bound on $1 - \gamma^k$, we get:-

$$1 - \gamma^k \leq \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)} \frac{\| x^{k+1} - x^* \|^2}{\| x^k - x^* \|^2}$$

$$\Rightarrow \sqrt{\left(\frac{\lambda_{min}(Q)}{\lambda_{max}(Q)}\right)(1 - \gamma^k)} \leq \frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|}$$

From lower bound on $(1 - \gamma^k)$,

$$\sqrt{\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}(1 - \gamma^k)} \geq \frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|} \quad (1)$$

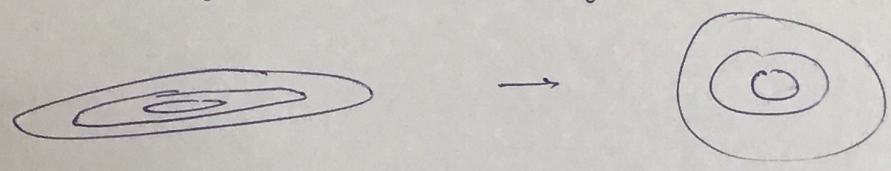
So, order of convergence is $p = 1$.

3) 4.3 As seen from the previous question, the order of convergence is dependence on eigen values of Hessian by.

$$\frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|} \leq \sqrt{\frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}} (1 - r(k))$$

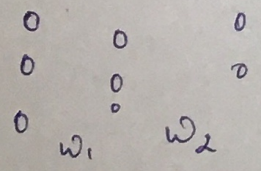
So, if $\lambda_{max}(Q) \simeq \lambda_{min}(Q)$, then we have faster convergence and smaller upper bound on $\frac{\| x^{k+1} - x^* \|}{\| x^k - x^* \|}$.

This is achieved by standardization:-
It makes $\lambda_{max}(Q)$ very close to $\lambda_{min}(Q)$.
and change in error in function surface is given by:-



6

Proof 6) 3 layer network



$\Rightarrow$ activation fn is f

output $= w_2 \, f(w_1 x)$

If $f(\cdot)$ is linear, i.e. $f(\vec{y}) = A\vec{y}$. Then

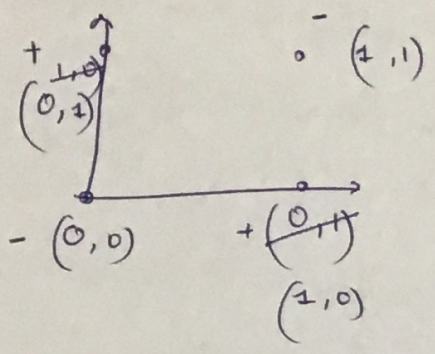
output $= w_2 \, A \, w_1 \vec{x}$

output $= W_x \vec{x}$; $w_2 A w_1 = W_x$

So output is linear in terms of input.

⇒ Unable to solve XOR Problem

with 2 binary input case, the labels look like :



⇒ Unable to shatter this surface with a linear classifier given by linear activation fn neuron.

Problem Ans 7.

⇒ $2^{|C|}$

7.1

(1) Let a set C, with cardinality $|C|$ be a set of boolean ~~function~~ points.

Then possible labellings over C is $2^{|C|}$.

So the hypothesis class contain atleast $2^{|C|}$ hypothesis fn.

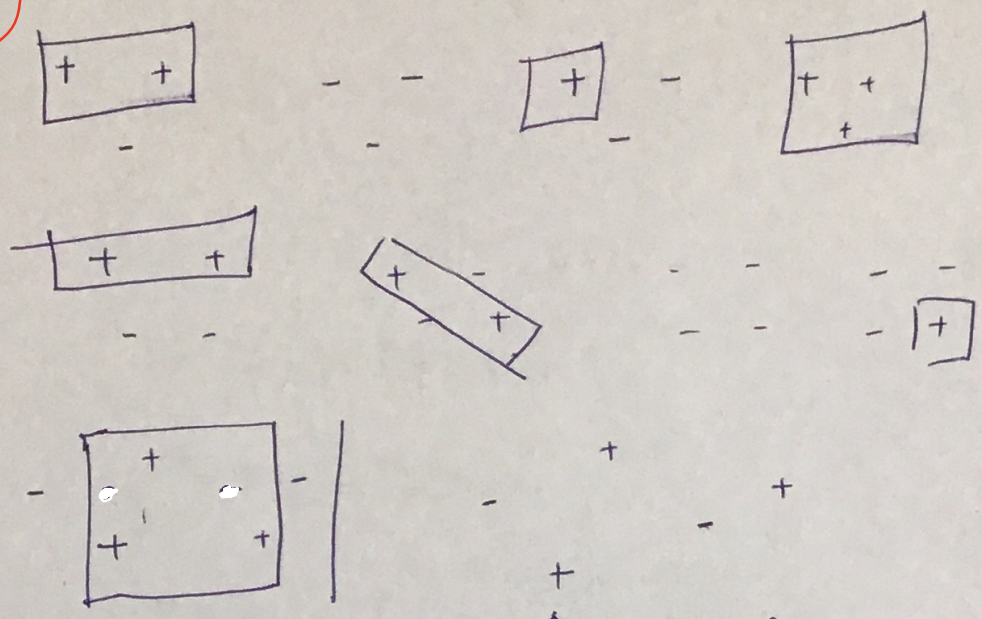
Therefore

$$|H| \geq 2^{|C|}$$

$$\log_2 |H| \geq \log_2 2^{|C|}$$

$$\log_2 |H| \geq |C|$$

3) 7.3



5 points

cannot shatter by rectangle.

Problem 5 6.1

1

(1) $E = E_0 + \frac{1}{2}(w - w^*)^T H (w - w^*)$

And update step:-

$$w^{\tau} = w^{\tau-1} - \rho \nabla E$$

and $\nabla E^{\tau-1} = H(w^{\tau-1} - w^*)$

$$w^{\tau} = w^{\tau-1} - \rho\left(H(w^{\tau-1} - w^*)\right)$$

Pre multiplying both side by u_j^T

$$u_j^T w^{\tau} = u_j^T w^{\tau-1} - \rho\, u_j^T H (w^{\tau-1} - w^*)$$

$$= w_j^{\tau-1} - \rho\, \eta_j u_j^T (w^{\tau-1} - w^*)$$

$$w_j^{\tau} = w_j^{\tau-1} - \rho\, \eta_j (w_j^{\tau-1} - w_j^*)$$

1.

T. P.

$$w_j^{(\tau)} = \left(1 - (1 - P x_j)^{\tau}\right) w_j^*$$

Proof by ~~recursion~~ induction :

$\tau = 1$

$\Rightarrow$ $\cancel{w_j^{(1)} = \left(1 - (1 - P x_j)^1\right)}$

$$w_j^{(1)} = w_j^{(0)} - P x_j \left(w_j^{(0)} - w_j^*\right)$$

As $w_j^{(0)} = 0$

Then $w_j^{(1)} = P x_j \, w_j^*$

$$= \left(1 - (1 - P x_j)\right) w_j^*$$

Assuming result is true for $\tau = N-1$, Then T. P. for $\tau = N$.

$$w_j^{(N)} = w_j^{(N-1)} - P x_j \left(w_j^{(N-1)} - w_j^*\right)$$

$$= w_j^{(N-1)} \left(1 - P x_j\right) + P x_j w_j^*$$

↓ Tome for this case is assumed.

$$= \left\{1 - (1 - P x_j)^{N-1}\right\} w_j^* \left(1 - P x_j\right) + P x_j w_j^*$$

$$w_j^{(N)} = \left\{1 - (1 - P x_j)^N\right\} w_j^*$$

5.02 2) Using result $w_j^{(\tau)} = \{1 - (1 - \rho x_j)^\tau\} w_j^*$

As $|1 - \rho x_j| < 1$

Then $|1 - \rho x_j|^\tau \to 0$, as $\tau \to \infty$.

Then $w_j^{(\tau)} \to w_j^*$

5.03 3) a) When $x \gg \frac{1}{\rho\tau} \Rightarrow \eta\rho\tau \gg 1$

And $|1 - \rho x_j| < 1 \Rightarrow$ means τ is a large number.

So as $\tau \to \infty$, $w_j^{(\tau)} \to w_j^*$ (followed from previous result).

b) T.P. $|w_j^{(\tau)}| \ll |w_j^*|$; when $\eta \ll (\rho\tau)^{-1}$

$\Rightarrow \eta\rho\tau \ll 1$, so τ does not need to be a large number

So using, $w_j^{(\tau)} = (1 - (1 - \rho x_j)^\tau) w_j^*$

approximating,

$(1 - \rho x_j)^\tau = 1 - \tau\rho x_j + O(\rho x_j^2)$

$\Rightarrow w_j^{(\tau)} = \{1 - (1 - \tau\rho x_j + O(\rho x_j^2))\} w_j^*$

$w_j^{(\tau)} = \tau\rho x_j w_j^*$

$w_j^{(\tau)} \ll w_j^*$.