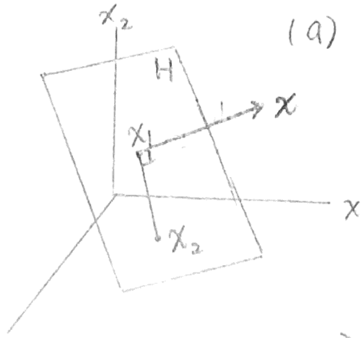
1.



(a) $g(\underline{x}) = w_0 + \underline{w}^T\underline{x}$

Assume x_1 and x_2 are the points on H

$g(x_1) = w_0 + \underline{w}^T\underline{x}_1 = 0$

$g(x_2) = w_0 + \underline{w}^T\underline{x}_2 = 0$

$g(\underline{x}_1) - g(\underline{x}_2) = \underline{w}^T(\underline{x}_1 - \underline{x}_2) = 0 \Rightarrow w^T \cdot (\underline{x}_1 - \underline{x}_2) = 0$

$\Rightarrow$ w is normal to H.

$(\underline{x}_1 - \underline{x}_2)$ is a vector on H.

(b) Assume there is a point x on $g(\underline{x}) > 0$ side, and $x = x_1 + a\underline{w}$ $(a > 0)$

the vector from x_1 to x is $(x - x_1) = a\underline{w}$

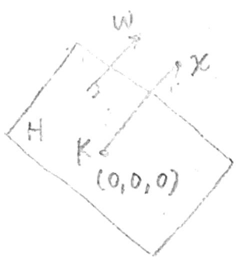
$(w, (x - x_1)) = w^T \cdot (x - x_1) = w^T \cdot a\underline{w} = a|w||w|\cos 0 = aw^2 > 0$

$\because$ the inner product of vector w and $(x - x_1)$ is bigger than 0

$\Rightarrow$ they point on the same side of H

$\Rightarrow$ w points to the positive side of H

(c)



$\because$ in "augmented feature space" $\Rightarrow$ H must pass through the origin K (0,0,0)

$r = \dfrac{g(x)}{\|w\|} = \dfrac{w^T x}{\|w\|}$, Assume there's a point x on the positive side of H.

$\underline{x} = a\underline{w} \xrightarrow{\text{take into } r} r = \dfrac{\underline{w}^T\underline{x}}{\|w\|} = \dfrac{\underline{w}^T(a\underline{w})}{\|w\|} = \dfrac{a\|w\|^2\cos 0}{\|w\|} = a\|w\| > 0$

$\because$ x is on the positive side of H(x)

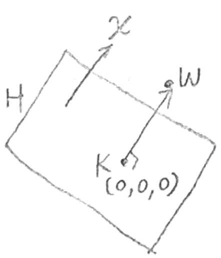
$\underline{x} = a\underline{w}$

when $a > 0$, $\underline{w}$ and $\underline{x}$ point to $r > 0$ side

$a < 0$, $a\underline{w}$ points to $r < 0$ side

$\Big\}$ $\underline{w}$ points to the $r > 0$ side of H.

(d)



$\because$ "augmented weight space" $\Rightarrow$ H must pass through the origin K (0,0,0)

$r = \dfrac{g(w)}{\|x\|} = \dfrac{w^T x}{\|x\|}$ Assume there's a point w on the positive side of H

$\underline{w} = a\underline{x} \xrightarrow{\text{take into } r} r = \dfrac{w^T x}{\|x\|} = \dfrac{a\|x\|^2\cos 0}{\|x\|} = a\|x\| > 0$

$\because$ w is on the positive side of H

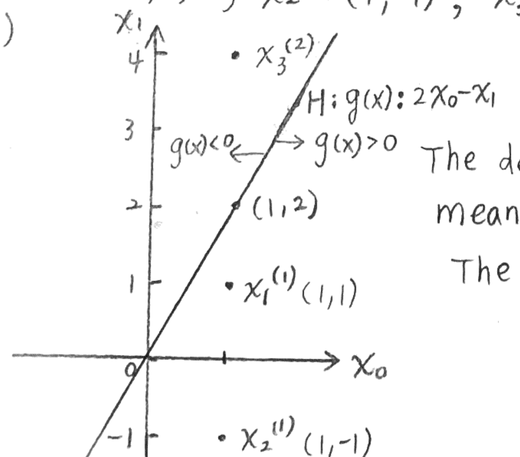
$\underline{w} = a\underline{x}$

when $a > 0$, $\underline{w}$ and $\underline{x}$ point to the positive side of H $(r > 0)$

$a < 0$, $\underline{w}$ points to the negative side of H. $(r < 0)$

2. $x_1^{(1)} = (1, 1)$, $x_2^{(1)} = (1, -1)$, $x_3^{(2)} = (1, 4)$

a)



Mean of class 1 : $(1, \frac{1+(-1)}{2}) = (1, 0)$

Mean of class 2 : $(1, 4)$

The decision boundry need to pass through (0,0) and the center of mean 1 and mean 2.

The side of class 1 is positive.

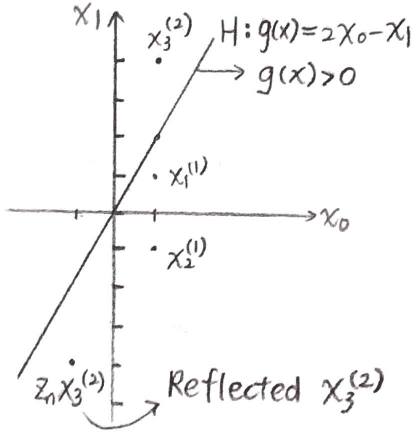
(b) Original : class 1 $\Rightarrow g(x) > 0$

class 2 $\Rightarrow g(x) < 0$

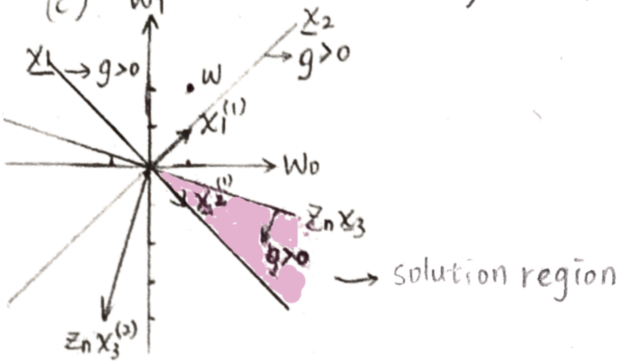
Reflected : $z_n^{(k)} \begin{cases} 1, & k=1 \\ -1, & k=2 \end{cases}$

$$w^T z_n^{(k)} x^{(k)} > 0$$

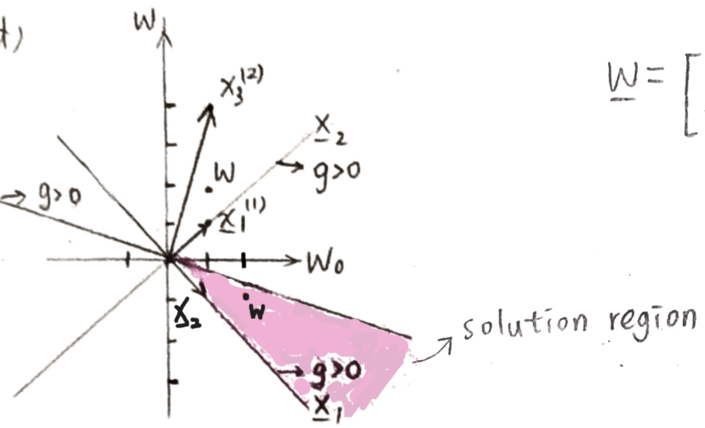
$\Rightarrow$ it still correctly classifies the points.



(c)



(d)



$w = \begin{bmatrix} 2 \\ -1 \end{bmatrix} \Rightarrow w$ is in the solution region.

3. (a) $f(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \\ \vdots \\ x_D(t) \end{bmatrix} \Rightarrow f(p) = \begin{bmatrix} x_1(p) \\ x_2(p) \\ \vdots \\ x_D(p) \end{bmatrix}$ $\quad p(\underline{x}) = \begin{bmatrix} p_1(\underline{x}) \\ p_2(\underline{x}) \\ \vdots \\ p_D(\underline{x}) \end{bmatrix}$

$$\nabla_{\underline{x}} f[p(\underline{x})] = \begin{bmatrix} \nabla_{\underline{x}} p(\underline{x}) \cdot \frac{d}{dp} x_1(p) \\ \nabla_{x} p(\underline{x}) \cdot \frac{d}{dp} x_2(p) \\ \vdots \\ \nabla_x p(\underline{x}) \cdot \frac{d}{dp} x_D(p) \end{bmatrix} = \begin{bmatrix} \frac{d}{dp} x_1(p) \\ \frac{d}{dp} x_2(p) \\ \vdots \\ \frac{d}{dp} x_D(p) \end{bmatrix} \nabla_{\underline{x}} p(\underline{x}) = \frac{d}{dp} \begin{bmatrix} x_1(p) \\ x_2(p) \\ \vdots \\ x_D(p) \end{bmatrix} \cdot \nabla_{\underline{x}} p(\underline{x}) = \left(\frac{d}{dp} f(p)\right) \nabla_{\underline{x}} p(\underline{x})$$

Q.E.D.

(b)

$$\frac{\partial}{\partial x}[x^t M x] = (M + M^t) x$$

$$\nabla_{\underline{x}} (\underline{x}^T \underline{x}) = \nabla_{\underline{x}} (x^T I x) = (I + I^t) \underline{x} = 2 I \underline{x} = 2 \underline{x}$$

(c)

$\underline{x} = \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ \underline{x}_D \end{bmatrix}$

$$\nabla_{\underline{x}} (\underline{x}^T \underline{x}) = \nabla_x \left([\underline{x}_1 \; \underline{x}_2 \; \underline{x}_3 \cdots \underline{x}_D] \begin{bmatrix} \underline{x}_1 \\ \underline{x}_2 \\ \vdots \\ x_D \end{bmatrix} \right) = \begin{bmatrix} \frac{\partial(\underline{x}_1^2 + \underline{x}_2^2 + \underline{x}_3^2 + \cdots + \underline{x}_D^2)}{\partial \underline{x}_1} \\ \frac{\partial(\underline{x}_1^2 + \underline{x}_2^2 + \cdots + \underline{x}_D^2)}{\partial \underline{x}_2} \\ \vdots \\ \frac{\partial(\underline{x}_1^2 + \underline{x}_2^2 + \cdots + \underline{x}_D^2)}{\partial \underline{x}_D} \end{bmatrix} = \begin{bmatrix} 2\underline{x}_1 \\ 2\underline{x}_2 \\ \vdots \\ 2\underline{x}_D \end{bmatrix} = 2\underline{x}$$

(d) $f(t) = t^3$, $p = \underline{x}^T\underline{x} \Rightarrow \nabla_x p(\underline{x}) = 2\underline{x}$

$$\nabla_{\underline{x}}[(\underline{x}^T\underline{x})^3] = \nabla_{\underline{x}} f[p(\underline{x})] = \frac{df(p)}{dp} \cdot \nabla_x p(\underline{x}) = 3p^2 \cdot 2\underline{x} = 3(\underline{x}^T\underline{x})^2 \cdot 2\underline{x} = 3\|x\|_2^4 \cdot 2\underline{x} = 6\|x\|_2^4 \cdot \underline{x}$$

4. (a) $W = \begin{bmatrix} W_1 \\ W_2 \\ \vdots \\ W_D \end{bmatrix}$ $\qquad \|W\|_2 = \sqrt{W_1^2 + W_2^2 + W_3^2 + \cdots + W_D^2}$

let $p = \underline{W}^T\underline{W} = \underline{W}^2$, $f(t) = t^{\frac{1}{2}}$

$$\nabla_{\underline{W}}\|\underline{W}\|_2 = \nabla_{\underline{W}} f[p(\underline{W})] = \left(\frac{d}{dp}f(p)\right)\nabla_{\underline{W}} p(\underline{W}) = \frac{1}{2}p^{-\frac{1}{2}} \cdot 2\underline{W} = \frac{1}{2}\cdot(\|\underline{W}\|_2^2)^{-\frac{1}{2}} \cdot 2\underline{W} = \frac{\underline{W}}{\|\underline{W}\|_2}$$

(b)

$$p(\underline{w}) = (\underline{M}\underline{w} - \underline{b})^T(\underline{M}\underline{w} - \underline{b}) = \underline{w}^T\underline{M}^T\underline{M}\underline{w} - \underline{w}^T\underline{M}^T\underline{b} - \underline{b}^T\underline{M}\underline{w} + \underline{b}^T\underline{b}$$

$$\nabla_{\underline{w}} P(w) = 2\underline{M}^T\underline{M}\underline{w} - M^Tb - M^Tb = 2\underline{M}^T\underline{M}\underline{w} - 2M^Tb$$

$$\nabla_{\underline{w}}\|\underline{M}\underline{w} - \underline{b}\| = \frac{1}{2}[(\underline{M}\underline{w} - \underline{b})^2]^{-\frac{1}{2}} \cdot (2\underline{M}^T\underline{M}\underline{w} - 2\underline{M}^T\underline{b})$$

$$= \frac{\underline{M}^T\underline{M}\underline{w} - \underline{M}^T\underline{b}}{\|\underline{M}\underline{w} - \underline{b}\|_2}$$

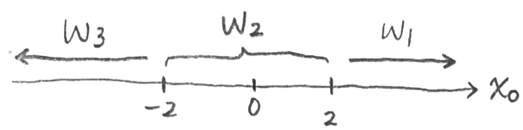
5. a) Show that total linear separability implies linear separability.

For points which are totally linear separable, for each class n there exists a hyperplane $H: g(x) = 0$ that separates the samples belonging to class n from the rest.

$$g(x) > 0 \quad \text{for } x \in w_n$$
$$g(x) < 0 \quad \text{for } x \notin w_n$$

b) Show that linear separability doesn't necessarily imply total linear separable.

Take a counterexample to illustrate:



$$\begin{cases} W_1: x_0 > 2 \\ W_2: -2 < x_0 < 2 \\ W_3: x_0 < -2 \end{cases}$$

W_1, W_2 and W_3 are linear separable, but they are not totally linear separable because there's no discriminant function $g(x)$ that can seperate W_2 samples from the rest.