H/W 6

2.1 a) $1(y \neq \text{sign}(f(x)) \leq \max\{0, 1 - yf(x)\}$

This can be solved considering two cases

Case 1 $y = \text{sign}(f(x))$

L.H.S = 0

R.H.S = $\max\{0, 1 - yf(x)\}$

Note that $yf(x)$ is +ve and the smallest value RHS can take is 0.

$\therefore$ L.H.S $\leq$ R.H.S.

Case 2 $y \neq \text{sign}(f(x))$

L.H.S = 1

R.H.S = $\max\{0, 1 - yf(x)\}$

Since $yf(x) \leq 0$, R.H.S will always be second term $1 - yf(x)$, which is greater than 1

$\therefore$ L.H.S $\leq$ R.H.S.

Hence proved.

b) hinge loss = $\max\{0, 1-m\}$

Note that a function $g(x) = 0$ is a straight line, and linear and hence convex.

$1-m$ is also a linear function in m and hence convex.

From the result of point wise maximum, we can say,

$\max\{0, 1-m\}$ is also convex since 0 and $1-m$ are convex.

c) $\max\{0, 1-y w^T x\}$

The expression $1-y w^T x$ is linear in w and hence $1-y w^T x$ is convex with respect to w.

∴ From previous analysis in part b,

$\max\{0, 1-y w^T x\}$ is also convex in w.

2.2) 1) $f(x) = \arg\max_{y \in Y} h(x,y)$

⇒ $f(x)$ is y that maximizes $h(x,y)$. For all other y, $h(x,y)$ is less.

∴ By def^n

$$h(x,y) \leq h(x, f(x)) \quad \text{— (I)}$$

2) $\Delta(y, f(x)) = \Delta(y, f(x))$

Adding eq (I) from part 1.

$$\Delta(y, f(x)) + h(x,y) \leq \Delta(y, f(x)) + h(x, f(x))$$

$$\Delta(y, f(x)) \leq \Delta(y, f(x)) + h(x, f(x)) - h(x,y)$$

From def^n of $f(x)$

$$\Delta(y, f(x)) \leq \max_{y' \in Y} [\Delta(y, y') + h(x, y') - h(x, y)]$$

$$\Delta(y, f(x)) \leq l(h, (x, y))$$

3. We have, $h_w(x, y) = \langle w, \psi(x, y) \rangle \mid w \in R^d$

and $l(h, (x, y)) = \max_{y' \in Y} [\Delta(y, y') + h(x, y') - h(x, y)]$

$$\therefore l(h_w, (x_i, y_i)) = \max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) \rangle - \langle w, \psi(x_i, y_i) \rangle]$$

Linear property of inner product

$$l(h_w, (x_i, y_i)) = \max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle]$$

4. a. $\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle$

Note that $\Delta(y_i, y)$ is a constant and $\langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle$ is a linear combination of values of w.

And hence this is a affine function of w.

b. Since $[\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i)\rangle]$ is an affine function in w. We can say it is convex. By property of pointwise maximum,

$\max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i)\rangle]$ is a convex function of w.

5. Since $l(h_w, (x_i, y_i)) \geq \Delta(y_i, f_w(x_i))$ and $l(h_w, (x_i, y_i))$ is convex in w as proved above.

$\therefore l(h_w(x_i, y_i))$ is a convex surrogate of $\Delta(y_i, f_w(x_i))$.

3. $l(h, (x, y)) = \max_{y' \in Y} [\Delta(y, y')) + h(x, y') - h(x, y)]$

There are two case:

Case 1: When $y = y'$

$\Delta(y, y') = 0$

$\Delta(y, y') + h(x, y') - h(x, y) = 0.$

Case 2: a) When $y \neq y'$ and $y = 1$ $(\Rightarrow y' = -1)$

$[\Delta(y, y') + h(x, y') - h(x, y)]$

$= 1 + -\frac{g(x)}{2} - \frac{g(x)}{2} = 1 - g(x)$

b) When $y \neq y'$ and $y = -1$ $(\Rightarrow y' = 1)$

$$[\Delta(y,y') + h(x,y') - h(x,y)]$$

$$= \left[1 + \frac{g(x)}{2} + \frac{g(x)}{2}\right] = 1 + g(x)$$

Generalizing case 2,

$$[\Delta(y,y') + h(x,y') - h(x,y)] = 1 - yg(x)$$

$$\therefore l(h,(x,y)) = \max\{0, 1 - yg(x)\}$$

4. 1. We proved that

$$l(h,(x_i,y_i)) = \max_{y' \in Y} [\Delta(y_i,y') + h(x_i,y') - h(x_i,y_i)]$$

But $m_{i,y}(h) = h(x_i,y_i) - h(x_i,y)$

$$\Rightarrow l(h,(x_i,y_i)) = \max_{y' \in Y} [\Delta(y_i,y') - m_{i,y}(h)]$$

2. To prove:

$$\max_{y \in Y} [(\Delta(y_i,y) - m_{i,y}(h))_+] = \max_{y \in Y} (\Delta(y_i,y) - m_{i,y}(h))$$

Note that when $y_i = y$

$$\Delta(y_i,y) - m_{i,y}(h) = \Delta(y_i,y) \geq 0$$

$\Rightarrow$ Hence is Across all values of $y \in Y$
$(\Delta(y_i, y) - m_{i,y}(h)) \geq 0$ for atleast one value.

$\Rightarrow$ No need to consider $(\,)_+$.

$$\therefore \max_{y \in Y} \left[\Delta'(y_i, y) - m_{i,y}(h)) \right]_+ = \max_{y \in Y} \left(\Delta(y_i, y) - m_{i,y}(h) \right)$$

3. When for all $y \neq y_i$

$$m_{i,y}(h) \geq \Delta(y_i, y)$$

$$\Rightarrow \left[\Delta(y_i, y) - m_{i,y}(h) \right] \leq 0.$$

and when $y = y_i$

$m_{i,y}(h) = 0$ and assumed $\Delta(y, y) = 0$

$$\Rightarrow \left[\Delta(y_i, y) - m_{i,y}(h) \right] = 0.$$

So, for all $y \in Y$

$$\max_{y \in Y} \left[\Delta(y_i, y) - m_{i,y}(h) \right] = 0.$$

5. 1. $J(w) = \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle]$

(a) We showed in question 2 that
$\max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle]$ is an affine function of w.

Σ of all affine function is also affine and hence
$\frac{1}{n} \sum_{i=1}^{n} \max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle]$

is affine and hence convex.

(b) We know that from properties
Every norm on R^n is convex and powers of convex is convex.
Hence $\|w\|^2$ is convex.

(c) $J(w)$ is the form sum of two convex function. We know from properties that sum of convex function is convex and hence $J(w)$ is convex.

differentiating.

3. Stochastic subgradient

$$\frac{\partial J(w)}{\partial w}_{stoch} = 2\lambda \|w\| + \times \max_{y \in Y} [\psi(x_i, y) - \psi(x_i,$$

4. Minibatch subgradient

$$\frac{\partial J(w)}{\partial w}_{minibatch} = 2\lambda \|w\| + \frac{1}{m} \sum_{j=i}^{i+m-1} \max_{y \in Y} [\psi(x_j, y) - \psi(x_j, y$$

2. $J(w) = \lambda \|w\|^2 + \frac{1}{n} \sum_{i=1}^{n} \hat{y}_i$ where

$$\hat{y}_i = \max_{y \in Y} [\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle]$$

$$\frac{\partial \hat{y}_i}{\partial w} = 0 + \langle w, \psi(x_i, \hat{y}_i) - \psi(x_i, y_i) \rangle$$

$$\therefore \frac{\partial J(w)}{\partial w} = 2\lambda w + \frac{1}{n} \sum_{i=1}^{n} \psi(x_i, \hat{y}_i) - \psi(x_i, y_i)$$

We take $\hat{y}$ because we have to take max by computing $\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle$, but n the max of $[\psi(x_i, y) - \psi(x_i, y_i)]$ which we get by differentiating.

3. stochastic subgradient

$$\frac{\partial J(w)}{\partial w}_{stoch} = 2\lambda \|w\| + \psi(x_i, \hat{y}) - \psi(x_i, y_i)$$

where $\hat{y} = \max_{y \in Y} \left[\Delta(y_i, y) + \langle w, \psi(x_i, y) - \psi(x_i, y_i) \rangle \right]$

4. similarly minibatch ~~gradi~~ subgradient will be

$$\frac{\partial J(w)}{\partial w}_{mini} = 2\lambda \|w\| + \frac{1}{m} \sum_{j=i}^{i+m-1} \psi(x_j, \hat{y}_j) - \psi(x_j, y_j)$$