Q 3.1

$$\underset{\alpha,\eta}{\text{prox}}[x] = \underset{v}{\text{argmin}} \frac{1}{2}\|v-x\|^2 + \alpha\eta(v)$$

for $\eta(v) = \lambda\|v\|_1$

$$\Rightarrow \underset{\|x\|_1,\alpha}{\text{prox}} = \underset{v}{\text{argmin}} \frac{1}{2}\|v-x\|^2 + \alpha\lambda\|v\|_1$$

$$= \underset{v_i}{\text{argmin}} \frac{1}{2}\|v_i-x_i\|^2 + \alpha\lambda|v_i| \quad \forall i$$

$$\overset{\forall i}{\Rightarrow} \underset{\|x\|_1,\alpha}{\text{prox}}[x_i] = \begin{cases} (v_i-x_i) + \alpha\lambda = 0 & \text{if } v_i > 0 \\ v_i - x_i - \alpha\lambda = 0 & \text{if } v_i < 0 \end{cases}$$

$$\Rightarrow \begin{cases} v_i = x_i - \alpha\lambda & \text{if } v_i > 0 \\ v_i = x_i + \alpha\lambda & \text{if } v_i < 0 \end{cases}$$

$$\Rightarrow \underset{\|x\|_1,\alpha}{\text{prox}}[x_i] = \begin{cases} x_i - \alpha\lambda & \text{if } x_i > \alpha\lambda \\ x_i + \alpha\lambda & \text{if } x_i < -\alpha\lambda \\ 0 & \text{if } -\alpha\lambda \le x_i \le \alpha\lambda \end{cases}$$

soft thresholding

Proximal gradient $\underset{x}{\operatorname{argmin}}\ f(x) + g(x)$

smooth non-smooth

$$\Rightarrow x^{t+1} = \underset{\alpha, r}{\text{Prox}}\left[x^t - \alpha_t \nabla f(x^t) \right]$$

Do gradient descent on smooth function, apply prox operator

3.1.2

$$\underset{w \in R^d}{\operatorname{argmin}}\ f(Xw - y) + \lambda \sum_{j=1}^{d} |w_j|$$

separable wrt w_j

$\Rightarrow$ do co-ordinate descent

$\Rightarrow$ 1) choose a random co-ordinate 'j'

$\rightarrow$ 2) Do gradient descent on that co-ordinate (using $\nabla_j f(x^t)$)

$\rightarrow$ apply co-ordinate - wise soft threshold operator

3.2 Essentially, repeat the single co-ordinate proof with blocks

3.3 Group - 21

$f(Xw - y)$ → multiclass logistic loss function

$W \in R^{k \times d}$

3.3.1 : $\underset{w}{\operatorname{argmin}}\ f(Xw - y) + \lambda \|w\|_F^2$

→ Do gradient descent $\sum_{c=1}^{k} \|w_c\|_2^2$

3.3.2 $\underset{w}{\operatorname{argmin}}\ f(Xw - y) + \lambda \|w\|_1$,

$\sum_{c=1}^{k} \|w_c\|_1$

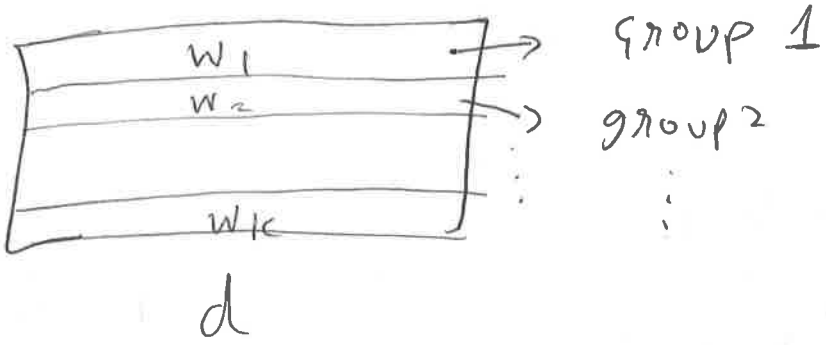
→ do proximal gradient $= \sum_{c=1}^{k} \left[\sum_{j=1}^{d} w_{cj} \right]$

3.3.3 Need to eliminate irrelevant features

w_{cj} → might be non-zero for some c

$\Rightarrow \forall c,\ w_{cj} = 0 \Rightarrow$ eliminate feature j

Do group-2,1 regularization

$W =$ 

- $\underset{W}{\operatorname{argmin}} f(Xw - y) + \lambda \sum_{g \in G} \|w_g\|$ → set of groups

→ use proximal gradient as before

with prox operator → $\frac{w_g}{\|w_g\|_2} \max\{0, \|w_g\|_2 - \lambda\}$

4.1.1 iterate averaging
→ maintain running average
$w_1, \frac{w_1 + w_2}{2}, \ldots$

4.1.2 average after $T/2$ iterations

4.1.3 optimize the optimization
→ step-size (decreasing, small constant)
→ Barzilai Bowdon step

4.2:

Use the $w^t = \beta^t v^t$ trick to ensure sparse updates

Projection onto τ- norm ball is expensive (dense)

$\Rightarrow$ maintain & update the L_2 norm of w^t in some way to ensure a quick (sparse) projection step.