2.1 $f(x) = \max_i f_i(x)$, $i = 1, \cdots, m$.

$f(x) \geq f_k(x)$, $k = 1, \cdots, m$.

Suppose that at a point x_0, $f(x_0) = f_k(x_0)$ and denote one of the subgradients of $f_k(x)$ at x_0 as g.

We have: $f(x) \geq f_k(x) \geq f_k(x_0) + g^T(x - x_0) = f(x_0) + g^T(x - x_0)$

So g also satisfies the definition of subgradient for $f(x)$.

2.2 $J(w) = \max\{0, 1 - yw^Tx\}$.

When $1 - yw^Tx \leq 0$, $J(w) = 0$, $\partial 0 = 0$.

When $1 - yw^Tx > 0$, $J(w) = 1 - yw^Tx$, $\partial(1 - yw^Tx) = -yx$

$$g = \begin{cases} 0, & 1 - yw^Tx \leq 0 \\ -yx, & 1 - yw^Tx > 0 \end{cases}$$ can be one of $J(w)$'s subgradients

3.1 If $\{x | w^Tx = 0\}$ is a seperating hyperplane for D.

then $y_i w^T x_i \geq 0$, $i = 1, \cdots, n$.

$\ell(\hat{y}, y) = \max\{0, -\hat{y}y\} = \max\{0, -yw^Tx\}$. so $\ell(\hat{y}_i, y_i) = 0$ and emprical loss is 0. Moreover, $\ell(\hat{y}, y) \geq 0$ so every seperating hyperplane is an empirical risk minimizer.

3.2 $\ell(\hat{y}, y) = \max\{0, -\hat{y}y\} = \max\{0, -yw^Tx\}$

$$g = \begin{cases} 0, & yw^Tx \geq 0 \\ -yx, & yw^Tx < 0 \end{cases}$$ can be one of the subgradients.

Only need to update w to $w + \eta y_i x_i$ when $y_i w^T x_i < 0$.

Choose a proper step size, we can implement stochastic subgradient descent now.

3.3. If w is initialized as 0, w shoube be linear combination of $\{x_1, \cdots, x_n\}$. Because it's added a scaled x_i in each update.

6.1 $J_i(w) = \frac{\lambda}{2}\|w\|_2^2 + \max\{0, 1 - y_i w^T x_i\}$

$$\nabla J_i(w) = \begin{cases} \lambda w, & 1 - y_i w^T x_i < 0 \\ \lambda w - y_i x_i, & 1 - y_i w^T x_i > 0 \end{cases}$$, when $y_i w^T x_i = 1$, it's undefined.

6.2 $\partial\left(\frac{\lambda}{2}\|w\|_2^2\right) = \lambda w$. $\partial \max\{0, 1 - y_i w^T x_i\} \ni \begin{cases} 0, & y_i w^T x_i \geq 1 \\ -y_i x_i, & \text{else} \end{cases}$.

$$g = \begin{cases} \lambda w, & y_i w^T x_i \geq 1 \\ \lambda w - y_i x_i, & \text{else} \end{cases} \in \partial J_i(w).$$

6.3 For step size $\eta_t = \frac{1}{\lambda t}$, $w^{(t+1)} = w^{(t)} - \eta_t g$

$$\eta_t g = \begin{cases} \frac{1}{t} w^{(t)} - \frac{1}{\lambda t} y_i x_i, & y_i w^{(t)} x_i < 1 \\ \frac{1}{t} w^{(t)}, & \text{else.} \end{cases}$$

$$w^{(t+1)} = \begin{cases} (1-\frac{1}{t}) w^{(t)} + \frac{1}{\lambda t} y_i x_i, & y_i w^{(t)} x_i < 1 \\ (1-\frac{1}{t}) w^{(t)}, & \text{else.} \end{cases}$$

6.5 $s_{t+1} = (1-\eta_t \lambda) s_t$ and let $w^{(t)} = s_t \cdot W_t$.

When $y_i w^{(t)} x_i \geq 1$, only need to update s_t.

When $y_i w^{(t)} x_i < 1$, $s_{t+1} W_{t+1} = (1-\eta_t \lambda) s_t \cdot W_t + \eta_t y_i x_i$.

So $W_{t+1} = W_t + \frac{1}{s_{t+1}} \eta_t y_i x_i$. should be the update rule of W_t.

Decomposing $w^{(t)}$ into s_t and W_t can significantly reduce the running time. Updating a number is much more timesaving than updating all parameters.