Solutions by **Andrew Lys** andrewlys(at)u.e.

1. **Kernelizing Gradient Descent.**

(a) First we compute the gradient of the loss function:

$$\begin{aligned}\nabla_w L_S(w) &= \frac{1}{m}\sum_{i=1}^{m}\nabla_w \ell(\langle w, \phi(x_i)\rangle; y_i) \\ &= \frac{1}{m}\sum_{i=1}^{m}\ell'(\langle w, \phi(x_i)\rangle; y_i)\nabla_w\langle w, \phi(x_i)\rangle \\ &= \frac{1}{m}\sum_{i=1}^{m}\ell'(\langle w, \phi(x_i)\rangle; y_i)\phi(x_i)\end{aligned}$$

Note that:

$$\begin{aligned}\langle w^{(t)}, \phi(x_i)\rangle &= \sum_{j=1}^{m}\alpha_j^{(t)}\langle \phi(x_j), \phi(x_i)\rangle \\ &= \sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i)\end{aligned}$$

So we have:

$$\nabla_w L_S(w^{(t)}) = \frac{1}{m}\sum_{i=1}^{m}\ell'\left(\sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i); y_i\right)\phi(x_i)$$

Therefore, we have:

$$\begin{aligned}w^{(t+1)} &= w^{(t)} - \eta\nabla_w L_S(w^{(t)}) \\ &= \sum_{i=1}^{m}\alpha_i^{(t)}\phi(x_i) - \eta\frac{1}{m}\sum_{i=1}^{m}\ell'\left(\sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i); y_i\right)\phi(x_i) \\ \sum_{i=1}^{m}\alpha_i^{(t+1)}\phi(x_i) &= \sum_{i=1}^{m}\left[\alpha_i^{(t)} - \eta\frac{1}{m}\ell'\left(\sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i); y_i\right)\right]\phi(x_i)\end{aligned}$$

Therefore, we have:

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} - \eta\frac{1}{m}\ell'\left(\sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i); y_i\right)$$

For each particular $\alpha_i^{(t)}$, we compute the kernel m times, m multiplication operations and $m-1$ addition operations. We also have the constant time operations of ℓ' and multiplying by $\frac{\eta}{m}$. Therefore, the total number of operations is $O(T_k \cdot m + m + (m-1)) = O(T_k \cdot m)$. We perform m of these operations, for each of the α's, so the total number of operations is $O(T_k \cdot m^2)$.

(b) The only term that is different here is the addition of $\frac{\lambda}{2}\|w\|_2^2$ to the loss function. We have the following gradient:

$$\nabla_w \|w\|_2^2 = 2w = 2\sum_{i=1}^{m}\alpha_i\phi(x_i)$$

Therefore, our new $\alpha_i^{(t+1)}$ is:

$$\alpha_i^{(t+1)} = \alpha_i^{(t)} - \eta\frac{1}{m}\ell'\left(\sum_{j=1}^{m}\alpha_j^{(t)}K(x_j, x_i); y_i\right) - \eta\lambda\alpha_i^{(t)}$$

$$= \alpha_i^{(t)}(1-\eta\lambda) - \eta\frac{1}{m}\ell'\left(\sum_{j=1}^m \alpha_j^{(t)} K(x_j, x_i); y_i\right)$$

(c) We investigate the $\|w\|_1$ term. We have the following gradient:

$$\nabla_w \|w\|_1 = \begin{bmatrix} \text{sign}(w_1) \\ \text{sign}(w_2) \\ \vdots \\ \text{sign}(w_d) \end{bmatrix} = \begin{bmatrix} \text{sign}(\sum_{i=1}^m \alpha_i \phi(x_i)_1) \\ \text{sign}(\sum_{i=1}^m \alpha_i \phi(x_i)_2) \\ \vdots \\ \text{sign}(\sum_{i=1}^m \alpha_i \phi(x_i)_d) \end{bmatrix}$$

It is not possible to write down this term in terms of the $K(x_i, x_j)$, since the sign function is not linear. Further, we have that

$$\nabla_w \|w\|_1 \in \{\pm 1\}^d$$

While it is not necessarily the case that the span of the data intersects this set. Therefore, it's possible that $\nabla_w \|w\|_1$ is linearly independent of the data, and so the iterates are also linearly independent of the data.

(d) Let $G = (K(x_i, x_j))_{i,j=1}^m$. Let $G[i]$ be the ith row of G, and let $G[:, j]$ be the jth column of G. Note that

$$\begin{aligned} \langle w(\alpha), \phi(x_i)\rangle &= \sum_{j=1}^m \alpha_j K(x_j, x_i) \\ &= \langle \alpha, G[:, i]\rangle \end{aligned}$$

Therefore, we have:

$$\begin{aligned} \nabla_\alpha \ell(\langle w(\alpha), \phi(x_i)\rangle; y_i) &= \nabla_\alpha \ell(\langle \alpha, G[:, i]\rangle; y_i) \\ &= \ell'(\langle \alpha, G[:, i]\rangle; y_i) G[:, i] \end{aligned}$$

And we have:

$$\begin{aligned} \nabla_\alpha L_S(w(\alpha)) &= \sum_{i=1}^m \ell'(\langle \alpha, G[:, i]\rangle; y_i) G[:, i] \\ &= G^\top (\ell'(\langle \alpha, G[:, i]\rangle; y_i))_{i=1}^m \end{aligned}$$

Therefore, we can write our α update as follows:

$$\alpha^{(t+1)} = \alpha^{(t)} - \eta\frac{1}{m} G^\top (\ell'(\langle \alpha^{(t)}, G[:, i]\rangle; y_i))_{i=1}^m$$

We can rewrite 1a. in terms of G as follows:

$$\begin{aligned} &\sum_{j=1}^m \alpha_j^{(t)} K(x_j, x_i) = \langle \alpha^{(t)}, G[:, i]\rangle \\ &\implies \alpha^{(t+1)} = \alpha^{(t)} - \frac{\eta}{m}(\ell'(\langle \alpha^{(t)}, G[:, i]\rangle; y_i))_{i=1}^m \end{aligned}$$

The place where this differs from the update we just computed is the in the multiplication by $G^\top$. Therefore, if we let

$$\phi(x_1) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$$

$$\phi(x_2) = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

We get

$$G = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}$$

So, we have:

$$G^\top (\ell'(\langle \alpha^{(t)}, G[:, i]\rangle; y_i))_{i=1}^2 \neq (\ell'(\langle \alpha^{(t)}, G[:, i]\rangle; y_i))_{i=1}^2$$

Since $G^\top$ is not the identity.

(e) We may write the gradient with respect to w as follows:

$$\begin{aligned}
\nabla_w \ell(\langle w^{(t)}, \phi(x_{i^{(t)}})\rangle; y_{i^{(t)}}) &= \ell'(\langle w^{(t)}, \phi(x_{i^{(t)}})\rangle; y_{i^{(t)}})\phi(x_{i^{(t)}}) \\
&= \ell'(\langle \sum_{j=1}^{m} \alpha_j^{(t)} \phi(x_j), \phi(x_{i^{(t)}})\rangle; y_{i^{(t)}})\phi(x_{i^{(t)}}) \\
&= \ell'(\sum_{j=1}^{m} \alpha_j^{(t)} K(x_j, x_{i^{(t)}}); y_{i^{(t)}})\phi(x_{i^{(t)}}) \\
&= \ell'(\langle \alpha^{(t)}, G[:, i^{(t)}]\rangle; y_{i^{(t)}})\phi(x_{i^{(t)}})
\end{aligned}$$

Therefore, the only coordinate that is changed in $\alpha^{(t+1)}$ is $\alpha_{i^{(t)}}^{(t+1)}$. Therefore, we have:

$$\begin{aligned}
\alpha_i^{(t+1)} &= \alpha_i^{(t)} && i \neq i^{(t)} \\
\alpha_i^{(t+1)} &= \alpha_{i^{(t)}}^{(t)} - \eta \ell'(\langle \alpha^{(t)}, G[:, i^{(t)}]\rangle; y_{i^{(t)}}) && i = i^{(t)}
\end{aligned}$$

In $G[:, i^{(t)}]$, we compute m kernel evaluations. Dotting this with $\alpha^{(t)}$ is $O(m)$ operations, and taking the loss and the other operations is constant. Therefore, the total number of operations is $O(T_k \cdot m)$.