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 $\ell'$ 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 $\alpha$'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 $i$th row of $G$, and let $G[:, j]$ be the $j$th 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 $\alpha$ 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 $O(T_k \cdot m)$.

2. **Implicit Regularization in Gradient Descent.**

(a) Suppose we have $w \in \text{lin}(\phi(x_1), \dots, \phi(x_m))$. such that

$$\Phi w = y$$

Since $w \in \text{lin}(\phi(x_1), \dots, \phi(x_m))$, we have that $w = \Phi^\top \alpha$ for some $\alpha \in \mathbb{R}^m$. Therefore, we have:

$$\begin{aligned}\Phi\Phi^\top \alpha &= y \\ \alpha &= (\Phi\Phi^\top)^{-1} y \\ \implies w &= \Phi^\top(\Phi\Phi^\top)^{-1} y\end{aligned}$$

Since $\Phi$ has full row rank, we have that $\Phi\Phi^\top$ is invertible. Therefore, $w$ indeed exists, and $(\Phi\Phi^\top)^{-1}y$ is unique. Additionally, since $\Phi$ has full row rank, we have that $\Phi^\top$ has full column rank, so $\Phi^\top$ is injective. Therefore, $w$ is unique, and $w^*$ exists. Finally, we have that $(\Phi\Phi^\top)^{-1}y \in \mathbb{R}^M$, so $w^* \in \text{lin}(\phi(x_1), \dots, \phi(x_m))$.

Now we show that this is the minimum norm solution, i.e.

$$w^* = \arg\min_{w \in \mathbb{R}^d, L_S(w)=0} \|w\|_2$$

Let $M = \text{lin}(\phi(x_1), \dots, \phi(x_m))$. Let $w \in \mathbb{R}^d$ such that $L_S(w) = 0$. Suppose that $w \notin M$. Let $P$ be the projection matrix onto $M$. i.e.

$$P = \Phi^\top(\Phi\Phi^\top)^{-1}\Phi$$

Notice that:

$$\begin{aligned}\|\Phi P w - y\|_2 &= \|\Phi\Phi^\top(\Phi\Phi^\top)^{-1}\Phi w - y\|_2 \\ &= \|\Phi w - y\|_2 = 0\end{aligned}$$

Since $Pw$ is in $M$ we have that $Pw = w^*$. We then have the following decomposition of $w$.

$$w = (w - Pw) + Pw = (w - Pw) + w^*$$

Where, $w - Pw$ is orthogonal $w^*$. Therefore, by the theorem of Pythagoras, we have:

$$\|w\|_2^2 = \|w - Pw\|_2^2 + \|w^*\|_2^2 > \|w^*\|_2^2$$

Therefore, if $w$ is a minimum norm solution, it must be in $M$, and as shown above, it must then be $w^*$.

(b) Let $M = \text{lin}(\phi(x_1), \dots, \phi(x_m))$. $w^{(0)} = 0 \in M$, so the base case is trivial. Suppose that $w \in M$. Then we have that $w = \Phi^\top \alpha$ for some $\alpha \in \mathbb{R}^m$. We have that:

$$\begin{aligned}w &= \Phi^\top \alpha \\ \Phi w &= \Phi\Phi^\top \alpha \\ \alpha &= (\Phi\Phi^\top)^{-1}\Phi w\end{aligned}$$

We know that $\Phi\Phi^\top$ is invertible, since $\Phi$ has full row rank. We focus on the gradient of $L_S(w)$:

$$\begin{aligned}
\nabla_w L_S(w) &= \frac{1}{m}\nabla_w \|\Phi w - y\|_2^2 \\
\nabla_w \|\Phi w - y\|_2^2 &= \nabla_w \|\Phi w\|_2^2 - \nabla_w 2\langle \Phi w, y\rangle + \nabla_w \|y\|_2^2 \\
&= \nabla_w \langle \Phi w, \Phi w\rangle - 2\nabla_w \langle \Phi\Phi^\top \alpha, y\rangle \\
&= \nabla_w w^\top \Phi^\top \Phi w - 2\nabla_w \langle \alpha, \Phi\Phi^\top y\rangle \\
&= 2\Phi^\top \Phi w - 2\nabla_w \alpha^\top \Phi\Phi^\top y \\
&= 2\Phi^\top \Phi\Phi^\top \alpha - 2\nabla_w w^\top \Phi^\top (\Phi\Phi^\top)^{-1} \Phi\Phi^\top y \\
&= 2\Phi^\top (\Phi\Phi^\top)\alpha - 2\Phi^\top y \\
&= \Phi^\top (2\Phi\Phi^\top \alpha - 2y) \in M \\
\implies \eta \nabla_w L_S(w) &\in M \\
\implies w - \eta\nabla_w L_S(w) &\in M
\end{aligned}$$

Therefore, $w^{(t)} \in M$ for all $t$.

We show directly that $w^{(t)} \to w^*$. Note that $\Phi^\top\Phi$ is a symmetric $d \times d$ matrix, so it has non-negative real eigenvalues. Since the row rank of $\Phi$ is $m$, we have $m$ non-zero eigenvalues, and the non-zero eigenvalues of $\Phi^\top\Phi$ are the eigenvalues of $\Phi\Phi^\top$. Let the maximum and minimum eigenvalues of $\Phi\Phi^\top$ be $\lambda$ and $\Lambda$. Note that for any scalar $\alpha$, the eigenvalues of

$$I - \alpha\Phi^\top\Phi$$

are between $1 - \alpha\Lambda$ and $1 - \alpha\lambda$. We prove this: Let $\mu$ be an eigenvalue of $I - \alpha\Phi^\top\Phi$, and let $v$ be an eigenvector. Then

$$\begin{aligned}
(I - \alpha\Phi^\top\Phi)v &= \mu v \\
(1-\mu)v &= \alpha\Phi^\top\Phi v
\end{aligned}$$

This implies that $\frac{1-\mu}{\alpha}$ is an eigenvalue of $\Phi^\top\Phi$, and so $\frac{1-\mu}{\alpha} \in [\lambda, \Lambda]$. Therefore,

$$\begin{aligned}
\lambda < \frac{1-\mu}{\alpha} &< \Lambda \\
\alpha\lambda < 1-\mu &< \alpha\Lambda \\
1 - \alpha\lambda > \mu &> 1 - \alpha\Lambda
\end{aligned}$$

If $\alpha > 0$ and

$$\begin{aligned}
\alpha\lambda > 1-\mu &> \alpha\Lambda \\
1 - \alpha\Lambda < \mu &< 1 - \alpha\lambda
\end{aligned}$$

if $\alpha < 0$. Either way, we have that $\mu$ is between $1 - \alpha\lambda$ and $1 - \alpha\Lambda$. Therefore, letting

$$\rho := \max\left(\left|1 - \frac{\eta}{2m}\lambda\right|, \left|1 - \frac{\eta}{2m}\Lambda\right|\right)$$

We have that

$$\|(I - \alpha\Phi^\top\Phi)v\|_2 \le \rho\|v\|_2$$

for all $v$. From above, we have:

$$\nabla L_S(w) = \frac{2}{m}\Phi^\top(\Phi w - y)$$

Let $w^{\mathrm{GD}}$ such that $\Phi w^{\mathrm{GD}} = y$. Then we have:

$$\nabla L_S(w^{\mathrm{GD}}) = 0$$

Therefore,

$$\|w^{(t+1)} - w^{\mathrm{GD}}\|_2 = \|w^{(t)} - \eta\nabla L_S(w^{(t)}) - w^{\mathrm{GD}}\|_2$$

$$
\begin{aligned}
&= \|w^{(t)} - w^{\mathrm{GD}} - \eta(\nabla L_S(w^{(t)}) - \nabla L_S(w^{\mathrm{GD}}))\|_2 \\
&= \|w^{(t)} - w^{\mathrm{GD}} - \eta\frac{2}{m}(\Phi^\top\Phi(w^{(t)} - w^{\mathrm{GD}}))\|_2 \\
&= \|(I - \eta\frac{2}{m}\Phi^\top\Phi)(w^{(t)} - w^{\mathrm{GD}})\|_2 \\
&\leq \rho\|w^{(t)} - w^{\mathrm{GD}}\|_2 \\
\implies &\leq \rho^{t+1}\|w^{(0)} - w^{\mathrm{GD}}\|_2 = \rho^{t+1}\|w^{\mathrm{GD}}\|_2
\end{aligned}
$$

Therefore, if $\rho < 1$, we have that $w^{(t)} \to w^{\mathrm{GD}}$. To make $\rho < 1$, we let

$$\eta = \frac{4m}{\lambda + \Lambda}$$

Then we have:

$$
\begin{aligned}
\rho &= \max\left(\left|1 - \frac{\eta}{2m}\lambda\right|, \left|1 - \frac{\eta}{2m}\Lambda\right|\right) \\
&= \max\left(\left|1 - \frac{2}{\lambda + \Lambda}\lambda\right|, \left|1 - \frac{2}{\lambda + \Lambda}\Lambda\right|\right) \\
&= \max\left(\left|\frac{\Lambda - \lambda}{\lambda + \Lambda}\right|, \left|\frac{\Lambda - \lambda}{\lambda + \Lambda}\right|\right) \\
&= \frac{\Lambda - \lambda}{\lambda + \Lambda} = \frac{1 - \frac{\lambda}{\Lambda}}{1 + \frac{\lambda}{\Lambda}} < 1
\end{aligned}
$$

Therefore, we have that $w^{(t)} \to w^{\mathrm{GD}}$. Thus, $w^{(t)}$ converges to a minimum norm solution. Further, since each $w^{(t)}$ is in $M$, and that $w^{(t)}$ is a convergent sequence, and that finite dimensional subspaces are always closed, $w^{(t)}$ converges to an element of $M$, so indeed $w^{\mathrm{GD}} \in M$. Since, $w^*$ is the unique zero error solution in $M$, we have $w^{\mathrm{GD}} = w^*$.