Quiz 3

1. (a) For constrained SVM

$$\min_{w,b,\varepsilon_i} \|w\|_2^2 + C/n \sum_{i=1}^{n} \varepsilon_i \quad ①$$
$$y_i(w^T x_i - b) \geq 1 - \varepsilon_i \quad \text{for } \forall i \quad ②$$
$$\varepsilon_i \geq 0 \quad \text{for } \forall i \quad ③$$

② is the same as $\varepsilon_i \geq 1 - y_i(w^T x_i - b)$

consider ② and ③ together, to min ε_i,

$\varepsilon_i \geq \max(0, 1 - y_i(w^T x_i - b)) = (1 - y_i(w^T x_i - b))_+$

plug it into ① get

$\min_{w,b} \|w\|_2^2 + C/n \sum_{i=1}^{n} (1 - y_i(w^T x_i - b))_+$, which is the above formula.

(b) For $w = \sum_j \alpha_j \phi(x_j) = \Phi^T \alpha$ for $\alpha \in R^n$ where $\Phi \in R^{n \times D}$ with $\phi(x_i)^T \in R^D$ as the ith row.

So $\phi(x_i)^T w = (\Phi \Phi^T \alpha)_i = (K\alpha)_i$, $K = \Phi\Phi^T$.

$\|w\|_2^2 = w^T w = \alpha^T \Phi \Phi^T \alpha$.

$$J(w) = C/n \sum_{i=1}^{n} (1 - y_i(\phi(x_i)^T w))_+ + \|w\|_2^2$$
$$= C/n \sum_{i=1}^{n} (1 - y_i(K\alpha)_i)_+ + \alpha^T K \alpha$$
$$= C\left(\frac{1}{n} \sum_{i=1}^{n} (1 - y_i(K\alpha)_i)_+ + \frac{1}{C} \alpha^T K \alpha\right)$$

So $\frac{1}{C} J(\alpha) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(K\alpha)_i)_+ + \lambda/2\, \alpha^T K \alpha$ where $\lambda/2 = 1/C$.

For $J(\alpha) = \frac{1}{n} \sum_{i=1}^{n} (1 - y_i(K\alpha)_i)_+ + \lambda/2\, \alpha^T K \alpha$,

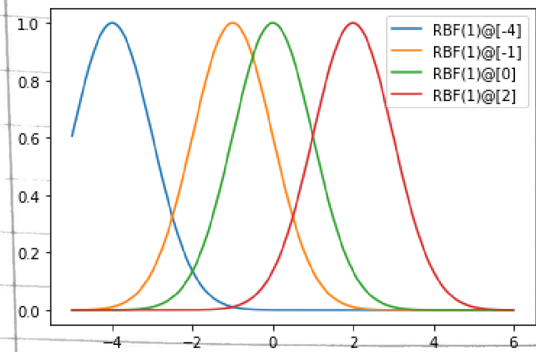
$$d/d\alpha\, J = \frac{1}{n} \sum_{i=1}^{n} t_i \cdot (-y_i \cdot k_i) + \lambda \cdot K \cdot \alpha$$

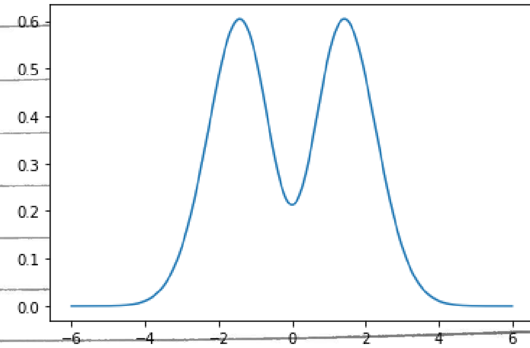
$$t_i = \begin{cases} 1 & \text{if } y_i(K\alpha)_i < 1 \\ 0 & \text{if } y_i(K\alpha)_i \geq 1 \end{cases}$$

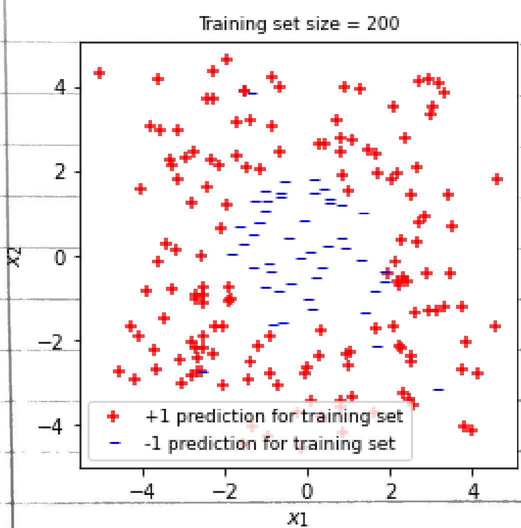
(c) ~~start from α.~~
~~for i in range(max_iteration):~~
~~if $y_i(K\alpha)_i < 1$:~~
~~gradient = $-y_i$~~

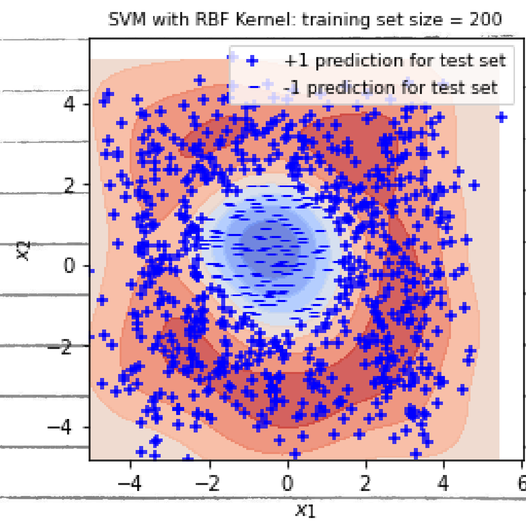
```
start from α.
for i in range(max_iteration):
        for j in range(x_length):
                if y_j (Kα)_j < 1:
                        gradient = -y_j · k_j + λ K α.
                else:
                        gradient = λ K α.
                α = α - t * gradient.
```

2,


RBF(1)@[-4]
RBF(1)@[-1]
RBF(1)@[0]
RBF(1)@[2]




Training set size = 200
+1 prediction for training set
-1 prediction for training set
x1
x2


SVM with RBF Kernel: training set size = 200
+1 prediction for test set
-1 prediction for test set
x1
x2

3. (a) safe arm $\sim Ber(\frac{1+\varepsilon}{2})$ $E(arm\ 1) = G_1/S_1$

risky arm $\sim Ber(\frac{1-\varepsilon}{2})$ $E(arm\ 2) = G_2/S_2$

$\begin{cases} S_1(1+\varepsilon) = 2G_1 \\ S_2(1-\varepsilon) = 2G_2 \end{cases}$ $\begin{cases} S_1 \cdot \varepsilon = 2G_1 - S_1 \\ S_2 \cdot \varepsilon = S_2 - 2G_2 \end{cases}$ combine

$(S_1+S_2)\varepsilon = 2G_1 - 2G_2 + S_2 - S_1$ where $\varepsilon > 0$ and $S_1 > 0$ and $S_2 > 0$.

So $2G_1 - 2G_2 + S_2 - S_1 > 0$.

If arm 2 is safe, we have the same expression as ~~abov~~ above except $\varepsilon < 0$. Then $2G_1 - 2G_2 + S_2 - S_1 < 0$.

Thus $2G_1 - 2G_2 + S_2 - S_1 > 0$, arm 1 is safe.

$2G_1 - 2G_2 + S_2 - S_1 < 0$, arm 2 is safe.

(b) For arm 1, set $[y_1, y_2, \cdots, y_n]$, $y_i \in \{0,1\}$, $r(arm\ 1) = G$.

For arm 2, set $[x_1, x_2, \cdots, x_n]$, $x_i \in \{0,1\}$, $x_i = |1 - y_i|$, $r(arm\ 2) = n - G$.

At a constant speed, since ε is small, we are put in higher risk of being stuck ~~at~~ in the risky arm. Thus can't go far.