Due: 24 Sept 2020

1. **Positive-definite.** Is the matrix $D = (yy^\intercal) \odot (XX^\intercal)$ positive-definite? Either prove this statement or give a counterexample.

Solution. First, note the diagonal entries of $XX^\intercal$ are $x_i^\intercal x_i = \|x_i\|^2$. Let $x_1 = 0$ and set all the other $x_i = (0, \ldots, 1, \ldots, 0)^\intercal$ where the 1 is in the ith position. Then the first row and column of $XX^\intercal$ are all zeros, all the other diagonal entries are 1, and the rest of the matrix is 0. When entrywise multiplied by $yy^\intercal$ (which is a symmetric matrix of all ± 1), the only entries that remain from $yy^\intercal$ are on the diagonal. Hence the remaining matrix is a diagonal matrix with a zero in the first diagonal spot and the rest of the diagonal entries are ± 1. Since the matrix is diagonal, the diagonal are its eigenvalues. The matrix here has a zero eigenvalue, which implies it cannot be positive-definite (not all the eigenvalues are positive). So the matrix is not always positive-definite.

2. **Dual problem.** Derive the dual problem for (52)-(54) in the notes (the case with soft margins).

Solution. The problem with soft margins is

$$
\begin{aligned}
f(w, b, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i \to \min_w \\
\text{subject to } y_i(w^\intercal x_i + b) \geq 1 - \xi_i, i = 1, \ldots, n \\
\xi_i \geq 0, i = 1, \ldots, n
\end{aligned}
$$

First, we need to find the Lagrangian and its gradient with respect to w:

$$
\begin{aligned}
L(w, b, \xi, \lambda) &= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i - \sum_{i=1}^n \lambda_i\big(y_i(w^\intercal x_i + b) - 1 + \xi_i\big) \\
&= \frac{1}{2}\|w\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \lambda_i\big(1 - \xi_i - y_i(w^\intercal x_i + b)\big).
\end{aligned}
$$

We need to find $\inf_{w,b,\xi} L(w, b, \xi, \lambda)$ in order to maximize over λ. To do this, first note that

$$
\begin{aligned}
\nabla_w L(w, b, \xi, \lambda) &= w - \sum_{i=1}^n \lambda_i y_i x_i \\
\frac{\partial L}{\partial b}(w, b, \xi, \lambda) &= -\sum_{i=1}^n \lambda_i y_i \\
\implies w^* &= \sum_{i=1}^n \lambda_i^* y_i x_i,\ (\lambda^*)^\intercal y = 0
\end{aligned}
$$

for optimal λ^*, w^*. Substituting this into the Lagrangian gives a function in λ and ξ:

$$
\begin{aligned}
q(\xi, \lambda) &= \frac{1}{2}\Big(\sum_{i=1}^n \lambda_i y_i x_i\Big)^\intercal\Big(\sum_{i=1}^n \lambda_i y_i x_i\Big) + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \lambda_i\Big[1 - \xi_i - y_i\Big(\Big(\sum_{j=1}^n \lambda_j y_j x_j^\intercal\Big)x_i + b\Big)\Big] \\
&= \frac{1}{2}\lambda^\intercal D\lambda + \sum_{i=1}^n (C - \lambda_i)\xi_i - \sum_{i=1}^n \lambda_i\Big[y_i\Big(\Big(\sum_{j=1}^n \lambda_j y_j x_j^\intercal\Big)x_i + b\Big) - 1\Big] \\
&= \frac{1}{2}\lambda^\intercal D\lambda - \lambda^\intercal D\lambda - by^\intercal\lambda + \sum_{i=1}^n \lambda_i + \sum_{i=1}^n (C - \lambda_i)\xi_i \\
&= -\frac{1}{2}\lambda^\intercal D\lambda + \sum_{i=1}^n \lambda_i + \sum_{i=1}^n (C - \lambda_i)\xi_i
\end{aligned}
$$

where D is defined as above (the other manipulations are akin to those we did in class). To find $\inf_\xi q(\xi, \lambda)$, we note that this clearly occurs when $\xi_i = 0$ for all i but that in that case we also must restrict $\lambda_i \leq C$, as otherwise the Lagrangian will no longer be bounded (the sum will diverge to $-\infty$). Incorporating that condition finally gives the dual problem for the problem with soft margins:

$$q(\lambda) = -\frac{1}{2}\lambda^\intercal D\lambda + \sum_{i=1}^{n} \lambda_i \to \max_{\lambda}$$
$$\text{subject to } 0 \le \lambda_i \le C,\ i = 1, \dots, n$$
$$\lambda^\intercal y = 0.$$

3. **Descent directions.** Let $(p^*, \lambda^*)^\intercal$ be a solution to the modified KKT system

$$\begin{bmatrix} \tilde{H} & A^\intercal \\ A & 0 \end{bmatrix} \begin{bmatrix} -p \\ \lambda \end{bmatrix} = \begin{bmatrix} \nabla f \\ 0 \end{bmatrix}$$

(see the end of Section 4 in the notes). Show that p^* is a descent direction, i.e., $\nabla f(x)^\intercal p^* < 0$—meaning that the motion along it for a sufficiently short distance will reduce the value of the objective function—provided that the columns of $A^\intercal$ are linearly independent and $n < d$. *Hint: first try to get it yourself, if you get stuck there is a paper provided on ELMS to look into.*

Solution. Mulitplying the system out into matrix form yields

$$\begin{cases} -\tilde{H}p + A^\intercal\lambda = \nabla f \\ -Ap = 0 \end{cases}.$$

The assumptions imply that there exists a solution to this system; denote this solution by (p^*, λ^*). Then left-multiplying the first equation by $p^{*\intercal}$ gives

$$\begin{aligned} &-p^{*\intercal}\tilde{H}p + p^{*\intercal}A^\intercal\lambda^* = p^{*\intercal}\nabla f \\ \implies &-p^{*\intercal}\tilde{H}p + (Ap)^\intercal\lambda^* = \nabla f^\intercal p^* \\ \implies &-p^{*\intercal}\tilde{H}p = \nabla f^\intercal p^* \end{aligned}$$

as $Ap^* = 0$ by the second equation in the system from before and $p^{*\intercal}\nabla f = \nabla f^\intercal p^*$ by symmetry. Finally, since $\tilde{H}$ is positive-definite, we have the desired result:

$$\nabla f^\intercal p = -p^{*\intercal}\tilde{H}p < 0$$

and hence p^* is a descent direction.

4. **Swiss roll.** Consider a Swiss roll dataset as shown in the notes. This dataset is generated by the provided Matlab code `stardata.m`. Design a nonlinear mapping to 2-dimensional or 3-dimensional feature space in which the blue and black sets are separable by a line or a plane. Visualize the data in the feature space so it is apparent that there exists such a separating line/plane. Submit a formula for your nonlinear map and your figure with the data mapped to the feature space. *You can also draw a linear divider but it is not required. There is a text file with the dataset provided as well.*

Solution. I created my nonlinear mapping by beginning with the mapping that worked in class ($f(x) = [\sin(3\phi(x)), \cos(3\phi(x))]^\intercal$ and modifying it in a way that seemed to work. After playing around with additional dimensions and various other tricks, I stumbled upon one trick that immediately worked: when computing the polar angle of the data points, subtract the radius of the point from the angle. This gives the function ϕ the form:

$$\phi(x) = \arctan 2(x_2, x_1) - \sqrt{x_1^2 + x_2^2}$$

where $\arctan 2(x, y)$ is the function returning the arctangent of x/y while choosing the angle correctly based on the signs of x and y. Plugging this into the original definition of f gives the following as my nonlinear map:

$$f(x) = [\sin(3(\arctan 2(x_2, x_1) - \sqrt{x_1^2 + x_2^2})), \cos(3(\arctan 2(x_2, x_1) - \sqrt{x_1^2 + x_2^2}))]$$

Subtracting off the radius yields the following division of the data:

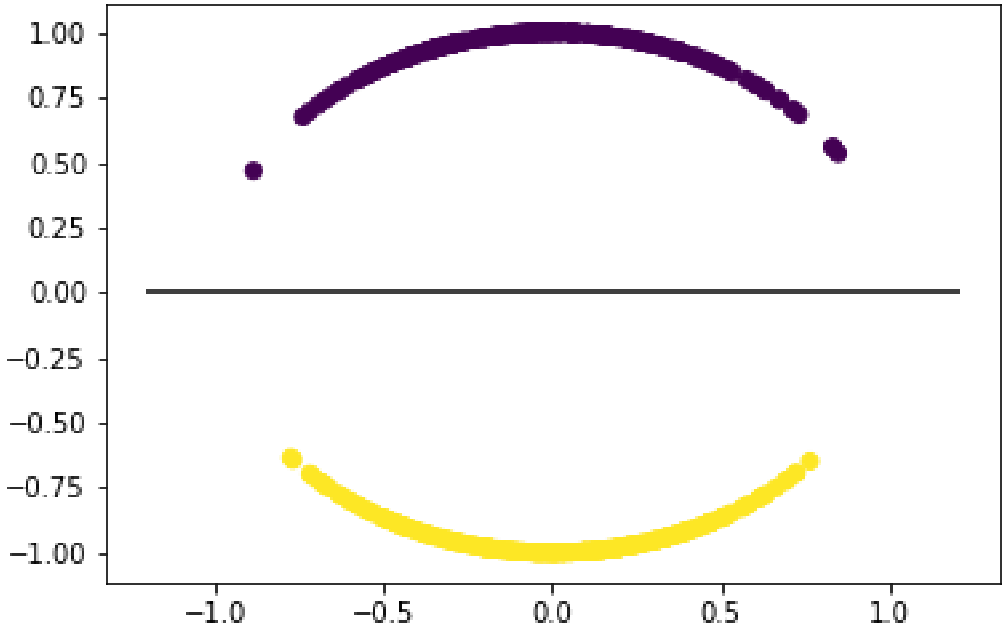

Figure 1: A depiction of feature space for the two clusters.