Objective : We consider a two-labelled (overlapping) dataset $\{(x_i, y_i)|i \in \{1, 2, .., n\}$ where $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ and want to find optimal margin linear and kernelized boundaries that separate the two labels.

Linear boundary case

Let $H(w, b) = 0$ the desired hyperplane. We define the "Geometrical Margin" M by

$$M \coloneqq \max_i M_i \text{ , where } M_i = dist(x_i, H) = \frac{w^T x_i + b}{||w||}.$$

$H(w, b)$ describes the hyperplane that lies among the data and has the largest possible distance from both classes. Note that due to the fact that data might overlap we also assign to each datapoint x_i a parameter ξ_i which will describe if the point is eventually on the correct side of the boundary or not.
The above is described via the optimization problem of "step 1" and can be solved by following the below steps :

Step 1:

$$\max_{\{w,b,\xi_i\}} M$$

subject to $y_i M_i \geq M(1 - \xi_i)$ and $\xi_i \geq 0$, for all $i \in \{1, 2, .., n\}$,
and $\sum_i \xi_i \leq K$, where K controls the number of points on the wrong side of the boundary.

Step 2: Turn the above problem into the below equivalent "min" problem :

$$\min_{\{w,b,\xi_i\}} \frac{1}{2}||w||^2 + C \sum_i \xi_i$$

subject to $y_i(w^T x_i + b) \geq (1 - \xi_i)$ and $\xi_i \geq 0$, for all $i \in \{1, 2, .., n\}$.

Remark By using the constraints on ξ_i, we can solve for ξ_i and replace it in the objective function of the min problem. Then the sum is with respect to the "Hinge loss" function for $p = 1$, given by

$$hinge(z) = max(0, 1 - z)^p$$

Step 3: Consider the equivalent "Primal" minimization problem using the Lagrangian approach, with Lagrange multipliers α_i, $i \in \{1, 2, .., n\}$.

Step 4: Verify that the "KKT" conditions hold and consider the equivalent "Dual" maximization problem.
After optimizing the Lagrangian with respect to w and b, we get closed form optimal formulas for w^* and b^*, and the problem takes the form :

$$\max_{\alpha_i} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle x_i \, , x_j \rangle \tag{1}$$

subject to $0 \leq \alpha_i \leq C$ for all $i \in \{1, 2, .., n\}$ and $\sum_i \alpha_i y_i = 0$.

Step 5: Use the SMO algorithm to solve the Dual problem.

Step 6: We compute the optimal margin hyperplane

$$H^*(x) = \sum_i \alpha_i^* y_i \langle x \, , x_i \rangle + b^* \tag{2}$$

Remark Note that the above boundary depends only to those points x_i for which $\alpha_i^* > 0$. In addition, from the "KKT" conditions we have that

$$\alpha_i^*(1 - y_i({w^*}^T x_i + b^*)) = 0 \text{ for all } i \in \{1, 2, .., n\}.$$

Consequently, all points that contribute to (2) should lie either on ${w^*}^T x_i + b^* = 1$ or on ${w^*}^T x_i + b^* = -1$. In other words all the points that "**support**" the hyperplane lie on the edge of the two classes.

Kernelized boundary case

In order to get a non-linear boundary $H^* = 0$ we must replace the inner product $\langle x_i \, , x_j \rangle$ in (1) and (2) above by $\langle \phi(x_i) \, , \phi(x_j) \rangle$ where ϕ is a non-linear function. However, the choice of the transforming function ϕ is not obvious and it would take a lot of time to find "well performing" choices. Instead, we use the "**Kernel Trick**" which allows us to replace the inner product $\langle x_i \, , x_j \rangle$ by well-known Kernel functions K which satisfy Mercer's theorem.

Popular examples include :

- the Polynomial kernel $K(x, z) = (x^T z + c)^d$, $d \geq 2$
- the Gaussian kernel $K(x, z) = \exp(-\frac{||x-z||^2}{2\sigma^2})$