Understanding α in SVM

In **Support Vector Machines (SVM)**, the **Lagrange multipliers** α_i play a crucial role in solving the optimization problem efficiently. They come from the **Lagrangian dual formulation** of the SVM problem, which allows us to express the problem in terms of dot products, making kernel methods possible.

1. Why Do We Need α?

We start with the **primal optimization problem**:

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

subject to:

$$y_i(w^T x_i + b) \geq 1, \quad \forall i$$

To solve this constrained optimization problem, we introduce **Lagrange multipliers** α_i (one for each training point) and construct the **Lagrangian function**:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \alpha_i [y_i(w^T x_i + b) - 1]$$

where $\alpha_i \geq 0$ are the **Lagrange multipliers**.

2. What Does α Represent?

- Each α_i determines how much influence a training point x_i has on defining the **optimal hyperplane**.
- If $\alpha_i > 0$, the corresponding point is a **support vector** (i.e., it lies on the margin or is misclassified in soft-margin SVM).
- If $\alpha_i = 0$, the corresponding point **does not contribute** to the hyperplane (i.e., it is correctly classified and far from the margin).

From the **Karush-Kuhn-Tucker (KKT) conditions**, a training point satisfies one of the following:

1. $\alpha_i = 0 \implies$ The point is correctly classified and lies far from the margin.
2. $\alpha_i > 0$ **but** $\alpha_i < C \implies$ The point is a **support vector** (on the margin).
3. $\alpha_i = C$ (in soft-margin SVM) $\implies$ The point is **misclassified or within the margin**.

Thus, only the **support vectors** (points with $\alpha_i > 0$) define the final decision boundary.

3. How α Helps in Finding w and b?

Using the KKT conditions, we derive:

$$w = \sum_{i=1}^{n} \alpha_i y_i x_i$$

This equation tells us:

- Instead of computing w directly, we express it as a **weighted sum** of training points.
- Only the **support vectors** contribute to w, because for non-support vectors, $\alpha_i = 0$.

For the bias term b, we use any support vector x_s with $\alpha_s > 0$:

$$b = y_s - w^T x_s$$