Assume we have samples $z_1 = (x_1, y_1), \ldots, z_n = (x_n, y_n)$ as well as a new sample z_{n+1}. The classifier trained on the data $z_1, \ldots, z_n$ is $f_{z_1,\ldots,z_n}$.
The error of this classifier is

$$\text{Error}(z_1, \ldots, z_n) = \mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}) = \mathbb{P}_{z_{n+1}}(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1})$$

and the *Average Generalization Error*

$$\text{A.G.E.} = \mathbb{E}\ \text{Error}(z_1, \ldots, z_n) = \mathbb{E}\mathbb{E}_{z_{n+1}} I(f_{z_1,\ldots,z_n}(x_{n+1}) \neq y_{n+1}).$$

Since $z_1, \ldots, z_n, z_{n+1}$ are i.i.d., in expectation training on $z_1, \ldots, z_i, \ldots, z_n$ and evaluating on z_{n+1} is the same as training on $z_1, \ldots, z_{n+1}, \ldots, z_n$ and evaluating on z_i. Hence, for any i,

$$\text{A.G.E.} = \mathbb{E}\mathbb{E}_{z_i} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)$$

and

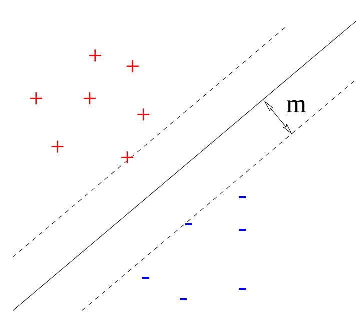
$$\text{A.G.E.} = \mathbb{E}\left[\underbrace{\frac{1}{n+1}\sum_{i=1}^{n+1} I(f_{z_1,\ldots,z_{n+1},\ldots,z_n}(x_i) \neq y_i)}_{\text{leave-one-out error}}\right].$$

Therefore, to obtain a bound on the generalization ability of an algorithm, it's enough to obtain a bound on its leave-one-out error. We now prove such a bound for SVMs. Recall that the solution of SVM is $\varphi = \sum_{i=1}^{n+1} \alpha_i^0 y_i x_i$.

Theorem 4.1.

$$L.O.O.E. \leq \frac{\min(\#\ \textit{support vect.}, D^2/m^2)}{n+1}$$

where D is the diameter of a ball containing all x_i, $i \leq n+1$ and m is the margin of an optimal hyperplane.



Remarks:

- dependence on sample size is $\frac{1}{n}$
- dependence on margin is $\frac{1}{m^2}$
- number of support vectors (sparse solution)

Lemma 4.1. *If x_i is a support vector and it is misclassified by leaving it out, then $\alpha_i^0 \geq \frac{1}{D^2}$.*

Given Lemma 4.1, we prove Theorem 4.1 as follows.

Proof. Clearly,

$$\text{L.O.O.E.} \leq \frac{\#\text{ support vect.}}{n+1}.$$

Indeed, if x_i is not a support vector, then removing it does not affect the solution. Using Lemma 4.1 above,

$$\sum_{i \in \text{supp.vect}} I(x_i \text{ is misclassified}) \leq \sum_{i \in \text{supp.vect}} \alpha_i^0 D^2 = D^2 \sum \alpha_i^0 = \frac{D^2}{m^2}.$$

In the last step we use the fact that $\sum \alpha_i^0 = \frac{1}{m^2}$. Indeed, since $|\varphi| = \frac{1}{m}$,

$$\begin{aligned}
\frac{1}{m^2} &= |\varphi|^2 = \varphi \cdot \varphi = \varphi \cdot \sum \alpha_i^0 y_i x_i \\
&= \sum \alpha_i^0 (y_i \varphi \cdot x_i) \\
&= \underbrace{\sum \alpha_i^0 (y_i(\varphi \cdot x_i + b) - 1)}_{0} + \sum \alpha_i^0 - b \underbrace{\sum \alpha_i^0 y_i}_{0} \\
&= \sum \alpha_i^0
\end{aligned}$$

□

We now prove Lemma 4.1.

Proof. Define

$$w(\alpha) = \sum \alpha_i - \frac{1}{2}\left(\sum \alpha_i y_i x_i\right)^2,$$

which we maximize under constraints

(1) $$\alpha_i \geq 0 \quad \text{and} \quad \sum y_i \alpha_i = 0.$$

Assume the following ordering on the support vectors when trained on $z_1, \ldots, z_n, z_{n+1}$:

$$\underbrace{\alpha_1^0, \ldots,}_{+} \underbrace{\ldots, \alpha_k^0, 0, \ldots, 0}_{-} = \alpha^0$$

where the first k points are the support vectors. Now, assume we leave out x_1 and make a mistake on it, and

$$\alpha_1 = 0. \tag{2}$$

Now we have

$$\underbrace{\overbrace{0, \ldots, 0,}^{\beta(i)=1} \overbrace{\alpha_1', \ldots,}^{\beta(i)=0}}_{+} \underbrace{\overbrace{\ldots, \alpha_\ell',}^{\beta(i)=1} \overbrace{0, \ldots, 0}^{\beta(i)=0}}_{-} = \alpha'$$

where $\beta \in \{0,1\}^n$.
Let $t > 0$ and suppose $\alpha' + t\beta$ satisfies optimization conditions (1). We know that

$$w(\alpha' + t\beta) \leq w(\alpha^0).$$

Hence,

$$w(\alpha^0) - w(\alpha') \geq w(\alpha + t\beta) - w(\alpha').$$

Moreover,

$$w(\alpha') = \sum \alpha_i' - \frac{1}{2}\left(\sum \alpha_i' y_i x_i\right)^2$$

and

$$\begin{aligned} w(\alpha' + t\beta) &= \sum \alpha_i' + t\sum \beta_i - \frac{1}{2}\left(\sum \alpha_i' y_i x_i + t\sum \beta_i y_i x_i\right)^2 \\ &= \sum \alpha_i' + t\sum \beta_i - \frac{1}{2}\left(\sum \alpha_i' y_i x_i\right)^2 - t\sum \alpha_i' y_i x_i \cdot \sum \beta_i y_i x_i - \frac{t^2}{2}\left(\sum \beta_i y_i x_i\right)^2. \end{aligned}$$

Hence,

$$\begin{aligned}
w(\alpha' + t\beta) - w(\alpha') &= t\sum \beta_i - t\underbrace{\sum \alpha'_i y_i x_i}_{\varphi'} \cdot \sum \beta_i y_i x_i - \frac{t^2}{2}\left(\sum \beta_i y_i x_i\right)^2 \\
&= t\sum \beta_i (1 - y_i \varphi' \cdot x_i) - \frac{t^2}{2}\left(\sum \beta_i y_i x_i\right)^2 \\
&= t\sum \beta_i (1 - y_i(\varphi' \cdot x_i + b)) + tb\underbrace{\sum \beta_i y_i}_{0} - \frac{t^2}{2}\left(\sum \beta_i y_i x_i\right)^2 \\
&= t(1 - y_1(\varphi' \cdot x_1 + b)) - \frac{t^2}{2}\left(\sum \beta_i y_i x_i\right)^2
\end{aligned}$$

Maximizing the above expression over t, we find

$$t = \frac{1 - y_1(\varphi' \cdot x_1 + b)}{\left(\sum \beta_i y_i x_i\right)^2} \geq 0.$$

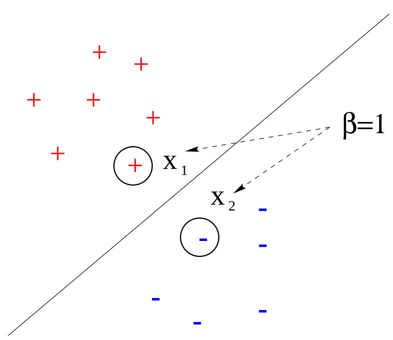
Substituting this t back into the expression,

$$w(\alpha' + t\beta) - w(\alpha') = \frac{\left(1 - y_1(\varphi' \cdot x_1 + b)\right)^2}{2\left(\sum \beta_i y_i x_i\right)^2}$$

Since x_1 is misclassified, $y_1(\varphi' \cdot x_1 + b) \leq 0$. Hence,

$$w(\alpha' + t\beta) - w(\alpha') \geq \frac{1}{2\left(\sum \beta_i y_i x_i\right)^2} \geq \frac{1}{2D^2}$$

because $|x_1 - x_2| \leq D$.



Now define γ as $\gamma(1) = \alpha_1^0$, $\gamma(i) = \alpha_i^0$ for $p \leq i \leq k$, and $\gamma(i) = 0$ otherwise, where

$$\underbrace{\alpha_1^0, \ldots,}_{+} \underbrace{\alpha_p^0, \ldots, \alpha_k^0, 0, \ldots, 0}_{-}.$$

We have

$$w(\alpha^0) - w(\alpha') \geq \frac{1}{2D^2}$$

and $\alpha^0 - \gamma$ satisfies constraint (2) and

$$w(\alpha^0 - \gamma) \leq w(\alpha').$$

$$\begin{aligned} w(\alpha^0) - w(\alpha') \leq w(\alpha^0) - w(\alpha^0 - \gamma) = \ & \text{... similarly to the previous proof} \\ = \frac{1}{2}\left(\sum \gamma_i y_i x_i\right)^2 = & \frac{(\alpha_1^0)^2}{2}\left(\sum \frac{\gamma_i}{\alpha_1^0} y_i x_i\right)^2 \\ = x_1 - \underbrace{\sum_{i=p}^{k} \frac{\gamma_i}{\alpha_1^0} x_i}_{\text{convex combination}} & \leq \frac{(\alpha_1^0)^2}{2} \cdot D^2 \end{aligned}$$

Hence,

$$\frac{1}{2D^2} \leq w(\alpha^0) - w(\alpha') \leq \frac{(\alpha_1^0)^2}{2} \cdot D^2$$

and so

$$\alpha_1^0 \geq \frac{1}{D^2}.$$

□