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

$$\text{Error}(z_1, \ldots, z_n) = E_{z_{n+1}} I(f_{z_1, \ldots, z_n}(x_{n+1}) \neq y_{n+1}) = P_{z_{n+1}} (f_{z_1, \ldots, z_n}(x_{n+1}) \neq y_{n+1})$$

and the Average Generalization Error

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

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 $i$,

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

and

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

leave-one-out error

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.**

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

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,

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

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

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

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

$$\frac{1}{m^2} = |\varphi|^2 = \varphi \cdot \varphi = \sum \alpha_i^0 y_i x_i$$

$$= \sum_0 \alpha_i^0 (y_i \varphi \cdot x_i)$$

$$= \sum_0 \alpha_i^0 (y_i (\varphi \cdot x_i + b) - 1) + \sum_0 \alpha_i^0 - b \sum_0 \alpha_i^0 y_i$$

$$= \sum_0 \alpha_i^0$$

We now prove Lemma 4.1. Let $u \bullet v = K(u, v)$ be the dot product of $u$ and $v$, and $\|u\| = (K(u, u))^{1/2}$ be the corresponding $L_2$ norm. Given $x_1, \ldots, x_{n+1} \in \mathbb{R}^d$ and $y_1, \ldots, y_{n+1} \in \{-1, +1\}$, recall that the primal problem of training a support vector classifier is $\arg\max_{\psi} \frac{1}{2} \|\psi\|^2$ subject to $y_i (\psi \cdot x_i + b) \geq 1$. Its dual problem is $\arg\max_{\alpha} \sum \alpha_i - \frac{1}{2} \|\sum \alpha_i y_i x_i\|^2$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$, and $\psi = \sum \alpha_i y_i x_i$. Since the Kuhn-Tucker condition can be satisfied, $\min_{\psi} \frac{1}{2} \psi \cdot \psi = \max_{\alpha} \sum \alpha_i - \frac{1}{2} \|\sum \alpha_i y_i x_i\|^2 = \frac{1}{2m^2}$, where $m$ is the margin of an optimal hyperplane.

Proof. Define $w(\alpha) = \sum \alpha_i - \frac{1}{2} \|\sum \alpha_i y_i x_i\|^2$. Let $\alpha^0 = \arg\max_{\alpha} w(\alpha)$ subject to $\alpha_i \geq 0$ and $\sum \alpha_i y_i = 0$. Let $\alpha' = \arg\max_{\alpha} w(\alpha)$ subject to $\alpha_p = 0$, $\alpha_i \geq 0$ for $i \neq p$ and $\sum \alpha_i y_i = 0$. In other words, $\alpha^0$ corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \ldots, n+1\}$ and $\alpha'$ corresponds to the support vector classifier trained from $\{(x_i, y_i) : i = 1, \ldots, p-1, p+1, \ldots, n+1\}$. Let $\gamma = \left( \begin{array}{c} 1 \\ \cdots \\ 0 \\ 1 \\ \cdots \\ 0 \end{array} \right)$. It follows that $w(\alpha^0 - \alpha_0^0 \cdot \gamma) \leq w(\alpha') \leq w(\alpha^0)$. (For the dual problem, $\alpha'$ maximizes $w(\alpha)$ with a constraint that $\alpha_p = 0$, thus $w(\alpha')$ is no less than $w(\alpha^0 - \alpha_0^0 \cdot \gamma)$, which is a special case that satisfies the constraints, including $\alpha_p = 0$. $\alpha^0$ maximizes $w(\alpha)$ with a constraint $\alpha_p \geq 0$, which raises the constraint $\alpha_p = 0$, thus $w(\alpha') \leq w(\alpha^0)$. For the primal problem, the training problem corresponding to $\alpha'$ has less samples $(x_i, y_i)$, where $i \neq p$, to separate with maximum margin, thus its margin $m(\alpha')$ is no less than the margin $m(\alpha^0)$.
and \( w(\alpha') \leq w(\alpha^0) \). On the other hand, the hyperplane determined by \( \alpha^0 - \alpha^0_p \cdot \gamma \) might not separate \((x_i, y_i)\) for \( i \neq p \) and corresponds to a equivalent or larger “margin” \( 1/\|\psi(\alpha^0 - \alpha^0_p \cdot \gamma)\| \) than \( m(\alpha') \).

Let us consider the inequality

\[
\max_t w(\alpha' + t \cdot \gamma) - w(\alpha') \leq w(\alpha^0) - w(\alpha') \leq w(\alpha^0) - w(\alpha^0 - \alpha^0_p \cdot \gamma).
\]

For the left hand side, we have

\[
w(\alpha' + t \gamma) = \sum \alpha'_i + t - \frac{1}{2} \sum \alpha'_i y_i x_i + t \cdot y_p x_p
\]

\[
= \sum \alpha'_i + t - \frac{1}{2} \parallel \sum \alpha'_i y_i x_i \parallel^2 - t \left( \sum \alpha'_i y_i x_i \right) \cdot (y_p x_p) - \frac{t^2}{2} \|y_p x_p\|^2
\]

\[
= w(\alpha') + t \cdot \left( 1 - y_p \cdot \left( \sum \alpha'_i y_i x_i \right) x_p \right) - \frac{t^2}{2} \|x_p\|^2
\]

and \( w(\alpha' + t \gamma) - w(\alpha') = t \cdot \left( 1 - y_p \cdot \psi' \cdot x_p \right) - \frac{t^2}{2} \|x_p\|^2 \). Maximizing the expression over \( t \), we find

\[
t = \left( 1 - y_p \cdot \psi' \cdot x_p \right)/\|x_p\|^2,
\]

and

\[
\max_t w(\alpha' + t \gamma) - w(\alpha') = \frac{1}{2} \left( 1 - y_p \cdot \psi' \cdot x_p \right)^2/\|x_p\|^2.
\]

For the right hand side,

\[
w(\alpha^0 - \alpha^0_p \cdot \gamma) = \sum \alpha^0_i - \alpha^0_p - \frac{1}{2} \sum \alpha^0_i y_i x_i - \alpha^0_p y_p x_p
\]

\[
= \sum \alpha^0_i - \alpha^0_p - \frac{1}{2} \|\psi_0\|^2 + \alpha^0_p y_p \psi_0 \cdot x_p - \frac{1}{2} \left( \alpha^0_p \right)^2 \|x_p\|^2
\]

\[
= w(\alpha_0) - \alpha^0_p (1 - y_p \cdot \psi_0 \cdot x_p) - \frac{1}{2} \left( \alpha^0_p \right)^2 \|x_p\|^2
\]

\[
= w(\alpha_0) - \frac{1}{2} \left( \alpha^0_p \right)^2 \|x_p\|^2.
\]

The last step above is due to the fact that \((x_p, y_p)\) is a support vector, and \( y_p \cdot \psi_0 \cdot x_p = 1 \). Thus \( w(\alpha^0) - w(\alpha^0 - \alpha^0_p \cdot \gamma) = \frac{1}{2} \left( \alpha^0_p \right)^2 \|x_p\|^2 \) and \( \frac{1}{2} \left( \frac{1 - y_p \cdot \psi' \cdot x_p}{\|x_p\|^2} \right)^2 \leq \frac{1}{2} \left( \alpha^0_p \right)^2 \|x_p\|^2 \). Thus

\[
\alpha^0_p \geq \frac{1 - y_p \cdot \psi' \cdot x_p}{\|x_p\|^2}
\]

\[
\geq \frac{1}{D^2}.
\]

The last step above is due to the fact that the support vector classifier associated with \( \psi' \) misclassifies \((x_p, y_p)\) according to assumption, and \( y_p \cdot \psi' \cdot x_p \leq 0 \), and the fact that \( \|x_p\| \leq D \).