Theorem 12. Suppose $||x|| \leq R$, $x \in \mathbb{R}^d$. Let $\mathcal{F}$ be a class of rules described by (5.4). Then

$$V-C \text{ dimension of } \mathcal{F} \leq \min \left\{ \frac{R^2}{\delta^2}, d \right\} + 1.$$  

Remark 9. What the above theorem says is that, for large enough margin $\delta$, the V-C dimension of $\mathcal{F}$ can be much smaller than $d + 1$. In high dimensional problems, this suggests a possibility of circumventing the curse of dimensionality.

5.3 Non-Separable Case

When the training data are not separable, some non-negative variables $\xi_i$'s are introduced to relax the separability condition:

$$\xi_i + y_i (\beta' x_i + \beta_0) \geq 1, \quad \xi_i \geq 0 \quad \text{for } i = 1, \ldots, n. \quad (5.5)$$

These $\xi_i$'s are often called slack variables in the optimization literature. Let $\xi = (\xi_1, \ldots, \xi_n)'$. Although introduction of the $\xi_i$'s makes it possible to relax the separability condition, if they are too large, then many data points could be incorrectly classified. If the $i$th data point is misclassified by the hyperplane $\beta' x + \beta_0 = 0$, that is, $y_i (\beta' x_i + \beta_0) \leq 0$, then $\xi_i \geq 1$. So, $\sum_{i=1}^n \xi_i$ provides an upper bound of the misclassification error of $\beta' x + \beta_0 = 0$. To maximize the margin and at the same time to minimize the bound, the SVM formulation for the separable case is modified to seek $(\beta_0, \beta, \xi)$ minimizing

$$\frac{1}{n} \sum_{i=1}^n \xi_i + \frac{\lambda}{2} ||\beta||^2$$

subject to (5.5). Here $\lambda$ is a positive tuning parameter that controls a trade-off between the error bound and the margin. By noting that $(\min \xi \text{ subject to } \xi \geq 0 \text{ and } \xi \geq a) = \max\{a, 0\} := a_+$ given $a$, it can be shown that the above modification equivalently finds $(\beta_0, \beta)$ that minimize

$$\frac{1}{n} \sum_{i=1}^n (1 - y_i (\beta' x_i + \beta_0))_+ + \frac{\lambda}{2} ||\beta||^2.$$

For a real-valued function $f(x) = \beta' x + \beta_0$ (instead of $f(x) = \text{sign}(\beta' x + \beta_0)$), $y_i (\beta' x_i + \beta_0)$ is called the functional margin of the individual point $(x_i, y_i)$ differently from the geometric margin in the separable case. The functional margin of $(x, y)$ is the product of a signed distance from $x$ to the hyperplane $\beta' x + \beta_0 = 0$ and $||\beta||$. If $y(\beta' x + \beta_0) > 0$,

$$y f(x) = ||\beta' x + \beta_0|| \times \text{distance}(x, \text{the hyperplane } \beta' x + \beta_0 = 0),$$

and otherwise

$$y f(x) = -||\beta' x + \beta_0|| \times \text{distance}(x, \text{the hyperplane } \beta' x + \beta_0 = 0).$$

The modified linear SVM formulation brings a new loss function to measure a goodness of fit of a classifier, which is given by

$$L(f(x_i), y_i) = (1 - y_i (\beta' x_i + \beta_0))_+ + (1 - y_i f(x_i))_+ = \xi_i.$$
It is known as the *hinge loss* as shown in Figure 5 together with the 0-1 loss (misclassification loss). Recall that the 0-1 loss is $L_{0-1}(f(x), y) = I(yf(x) \leq 0)$ for a real-valued discriminant function that induces a classifier through $\text{sign}(f(x))$. The hinge loss is a convex upper bound of the 0-1 loss and is monotonically decreasing in $yf(x) = y(\beta'x + \beta_0)$, the functional margin. The hinge loss makes the SVM computationally more attractive than direct minimization of empirical error rate.

Figure 5: Comparison of the hinge and 0 – 1 loss

In the non-separable case, the geometric interpretation of $2/\|\beta\|$ as the separation margin between two classes no longer holds although $2/\|\beta\|$ is often treated as a ‘soft’ margin analogous to the ‘hard’ margin in the separable case. Rather, $\|\beta\|^2$ can be regarded as a penalty imposed on the linear discriminant function $f$.

Hence, the SVM procedure can be cast in the regularization framework where a function estimation method is formulated as an optimization problem of finding $f$

$$
\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(f(x_i), y_i) + \lambda J(f).
$$

Here $L(f(x), y)$ is a loss function, $J(f)$ is a regularizer or a penalty imposed on $f$, and $\lambda > 0$ is a tuning parameter which controls the trade-off between data fit and the complexity of $f$. There are numerous examples of regularization procedures in statistics. For example, consider the multiple linear regression with $\mathcal{F} = \{f(x) = \beta'x + \beta_0 : \beta \in \mathbb{R}^d, \beta_0 \in \mathbb{R}\}$ and the squared error loss $L(f(x), y) = (y - f(x))^2$. $J(f) = \|\beta\|^2$ defines the ridge regression procedure while the LASSO takes $J(f) = \sum_{j=1}^{d} |\beta_j|$ as a penalty for a sparse linear model. Note that the SVM uses the ridge-like $\ell_2$ norm of $\beta$ as a penalty, $J(f) = \|\beta\|^2$. In other words, the SVM can be viewed as a procedure for penalized risk minimization with respect to the hinge loss. This viewpoint also connects the V-C dimension, the theoretical notion of capacity of $\mathcal{F}$ to more classical measure of complexity of a ‘model’ (or classifier) space implicitly. So, restriction of the model space by the size of $J(f)$ can be taken as a way of controlling the V-C dimension.
5.4 Constrained Optimization

For simplicity of discussion of optimization for the SVM, the separable case is considered first. The optimal hyperplane is determined by solving the following problem:

\[
\text{minimize } \frac{1}{2}||\beta||^2 \text{ subject to } y_i(\beta^T x_i + \beta_0) \geq 1, \ i = 1, \ldots, n.
\]

So, the primal problem has the objective function

\[
l_P(\beta, \beta_0) := \frac{1}{2} \beta^T \beta,
\]

which is free from \(\beta_0\), and \(n\) inequality constraints

\[
h_i(\beta, \beta_0) := 1 - y_i(\beta^T x_i + \beta_0) \leq 0, \ i = 1, \ldots, n.
\]

To handle the inequality constraints, the Lagrange multipliers or dual variables (\(\alpha_i\) for \(h_i(\beta, \beta_0) \leq 0\)) are introduced and the dual objective function is formed:

\[
l_D(\beta, \beta_0, \alpha) := l_P(\beta, \beta_0) + \sum_{i=1}^{n} \alpha_i h_i(\beta, \beta_0) = \frac{1}{2} \beta^T \beta + \sum_{i=1}^{n} \alpha_i (1 - y_i(\beta^T x_i + \beta_0)).
\]

Let \(\alpha = (\alpha_1, \ldots, \alpha_n)^T\). Then the dual problem becomes

\[
\text{maximize } l_D(\beta, \beta_0, \alpha)
\]

subject to \(\alpha_i \geq 0\) for all \(i = 1, \ldots, n\) and \(\nabla_{(\beta,\beta_0)} l_D(\beta, \beta_0, \alpha) = 0\). The equality constraints give

\[
\frac{\partial l_D}{\partial \beta} = \beta - \sum_{i=1}^{n} \alpha_i y_i x_i = 0 \Rightarrow \beta = \sum_{i=1}^{n} \alpha_i y_i x_i,
\]

\[
\frac{\partial l_D}{\partial \beta_0} = -\sum_{i=1}^{n} \alpha_i y_i = 0 \Rightarrow \sum_{i:y_i=1} \alpha_i = \sum_{i:y_i=-1} \alpha_i.
\]

Simplifying the objective function, we get

\[
l_D(\beta, \beta_0, \alpha) = \frac{1}{2} \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right)' \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right) + \sum_{i=1}^{n} \alpha_i - \left( \sum_{i=1}^{n} \alpha_i y_i x_i' \right) \left( \sum_{i=1}^{n} \alpha_i y_i x_i \right)
\]

\[
= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j
\]

\[
= l_D(\alpha) \quad \text{(say)}.
\]

Thus the dual SVM problem is

\[
\text{maximize } l_D(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j x_i' x_j
\]

subject to

\[
\alpha_i \geq 0 \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0.
\]