Let $Y = \text{diag}(y_1, \ldots, y_n)$, $K = [x_i' x_j]_{n\times n}$ (Gram matrix) and $\mathbf{1} = (1, \ldots, 1)'_{n\times 1}$. Then the dual problem is

$$\underset{\alpha}{\text{maximize}}\ l_D(\alpha) = \mathbf{1}'\alpha - \frac{1}{2}\alpha' Y K Y \alpha$$

subject to

$$\alpha \geq 0 \text{ and } \mathbf{1}' Y \alpha = 0,$$

which is a quadratic programming problem with a non-negative definite matrix YKY.

According to the constrained optimization theory, necessary and sufficient conditions (known as the Karush-Kuhn-Tucker (KKT) conditions) for the optimality of $\hat{\beta}_0$, $\hat{\beta}$, and $\hat{\alpha}$ are

1. Primal feasibility:
$$y_i(\hat{\beta}' x_i + \hat{\beta}_0) \geq 1 \quad \text{for all } i = 1, \ldots, n. \tag{5.6}$$

2. Dual feasibility:
$$\begin{aligned} \hat{\alpha}_i &\geq 0 \quad \text{for all } i = 1, \ldots, n, \\ \sum_{i=1}^{n} \hat{\alpha}_i y_i &= 0, \\ \hat{\beta} &= \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i \end{aligned} \tag{5.7}$$

3. Complementarity conditions:
$$\hat{\alpha}_i \Big(1 - y_i(\hat{\beta}' x_i + \hat{\beta}_0)\Big) = 0 \quad \text{for all } i = 1, \ldots, n. \tag{5.8}$$

Let $\hat{\alpha}$ be the solution of the dual problem. Then $\hat{\beta} = \sum \hat{\alpha}_i y_i x_i$. Once $\hat{\beta}$ is determined, $\hat{\beta}_0$ is obtained by using the complementarity condition for any data point with $\hat{\alpha}_i > 0$:

$$\hat{\alpha}_i \Big(1 - y_i(\hat{\beta}' x_i + \hat{\beta}_0)\Big) = 0.$$

Thus, we have $y_i(\hat{\beta}' x_i + \hat{\beta}_0) = 1$, which gives $\hat{\beta}_0 = y_i - \hat{\beta}' x_i$. Note that there are only a few data points with a positive Lagrange multiplier, and they determine the optimal separating hyperplane, which is given by

$$\hat{\beta}' x + \hat{\beta}_0 = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i' x + \hat{\beta}_0 = \sum_{i:\hat{\alpha}_i > 0} \hat{\alpha}_i y_i x_i' x + \hat{\beta}_0 = 0.$$

Such data points $\{(x_i, y_i) : \hat{\alpha}_i > 0\}$ are called the *support vectors*. See Figure 6. The encircled points are the support vectors with $\hat{\alpha}_i > 0$ and $y_i(\hat{\beta}' x_i + \hat{\beta}_0) = 1$. For those points outside the margin, $y_i(\hat{\beta}' x_i + \hat{\beta}_0) > 1$, which implies $\hat{\alpha}_i = 0$. The SVM solution is *sparse* in terms of the data points. The sparsity is due to the singularity of the hinge loss at 1.

Consider a much smaller problem of finding the optimal separating hyperplane with only the support vectors, namely, the points (x_i, y_i) such that $\hat{\alpha}_i > 0$. Let $S = \{i : \hat{\alpha}_i > 0\}$. Then, it is very trivial to see that the new estimates $\hat{\beta}_0^S$, $\hat{\beta}^S$, and $\hat{\alpha}_i^S$ for $i \in S$ will satisfy

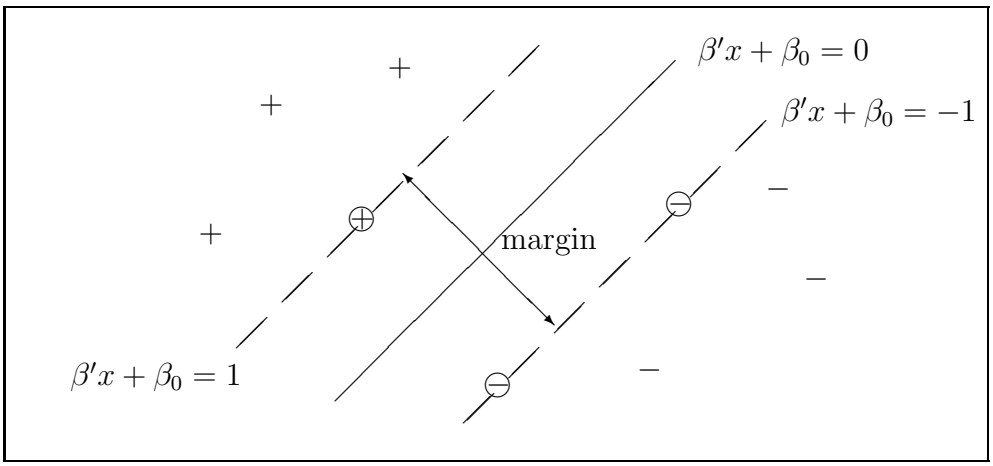

Figure 6: Support vectors

the original optimality conditions (5.6-5.8), which means that they are also solutions to the complete data problem. Hence removing any support vectors does not change the solution.

For classification of a new point x, we use

$$\hat{f}(x) = \text{sign}(\hat{\beta}'x + \hat{\beta}_0).$$

Note that this final form of $\hat{\beta}'x + \hat{\beta}_0$ does not depend on the dimensionality of x explicitly but on the inner products of x_i and x, and so does the dual problem. This fact enables us to construct hyperplanes even in infinite-dimensional Hilbert spaces (p.406, Vapnik 1998).

By a similar derivation as in the separable case, it can be shown that for the soft-margin SVM, the dual problem is

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

subject to

$$0 \leq \alpha_i \leq 1 \text{ for all } i = 1, \ldots, n \text{ and } \sum_{i=1}^{n} \alpha_i y_i = 0.$$

Also, the linear discriminant function is given by

$$\hat{f}(x) = \hat{\beta}'x + \hat{\beta}_0 = \sum_{i=1}^{n} \hat{\alpha}_i y_i x_i' x + \hat{\beta}_0.$$

5.5 Constrained Optimization Theory

This section gives a brief overview of the constrained optimization theory, applications of which abound in practice.

Primal problem:

$$\underset{x\in\mathbb{R}^d}{\text{minimize}}\ l_P(x) \quad \text{subject to } h_i(x) \le 0 \text{ for all } i = 1, \ldots, m \tag{5.9}$$

where l_P and $h_i : \mathbb{R}^d \to \mathbb{R}$ are differentiable functions. $l_P(\cdot)$ is a primal objective function and $\{h_i(\cdot)\}$ are m inequality constraints. x is called the *primal variable(s)*.

Dual problem: Introduce $\alpha_i \ge 0$ (Lagrange multiplier or dual variable) for $h_i(x) \le 0$, $i = 1, \ldots, m$. Let $\alpha = (\alpha_1, \ldots, \alpha_m)'$ and define the dual objective function

$$l_D(x, \alpha) := l_P(x) + \sum_{i=1}^{m} \alpha_i h_i(x). \tag{5.10}$$

Then, the dual problem is stated as

$$\underset{\alpha}{\text{maximize}}\ \underset{x}{\text{minimize}}\ l_D(x, \alpha) \text{ subject to } \alpha \ge 0$$

or equivalently

$$\underset{x,\alpha}{\text{maximize}}\ l_D(x, \alpha) \quad \text{subject to } \nabla_x l_D(x, \alpha) = \nabla l_P(x) + \sum_{i=1}^{m} \alpha_i \nabla h_i(x) = 0 \text{ and } \alpha \ge 0, \tag{5.11}$$

where $\nabla_x l = (\partial l/\partial x_1, \ldots, \partial l/\partial x_n)'$ and $\alpha \ge 0$ means $\alpha_i \ge 0$ for all $i = 1, \ldots, m$.

Definition 4 (primal/dual feasibility)**.** We say that $x \in \mathbb{R}^d$ is *primal feasible* if

$$h_i(x) \le 0 \quad \text{for all } i = 1, \ldots, m; \tag{5.12}$$

and (x, α) is *dual feasible* if

$$\nabla_x l_D(x, \alpha) = \nabla l_P(x) + \sum_{i=1}^{m} \alpha_i \nabla h_i(x) = 0 \text{ and } \alpha \ge 0. \tag{5.13}$$

Theorem 13 (Weak duality theorem)**.** *For a primal feasible x^P and a dual feasible (x^D, α^D),*

$$l_P(x^P) \ge l_D(x^D, \alpha^D) = l_P(x^D) + \sum_{i=1}^{m} \alpha_I^D h_i(x^D) \tag{5.14}$$

where l_P, h_i are assumed to be convex at x^D.

Proof.

$$l_P \text{ is convex at } x^D \Rightarrow l_P(x^P) - l_P(x^D) \geq \nabla l_P(x^D)(x^P - x^D) \tag{5.15}$$
$$\text{and } h_i \text{ is convex at } x^D \Rightarrow h_i(x^P) - h_i(x^D) \geq \nabla h_i(x^D)(x^P - x^D). \tag{5.16}$$

Now,

$$\begin{aligned}
l_P(x^P) - l_D(x^D, \alpha^D) &= l_P(x^P) - l_P(x^D) - \sum_{i=1}^{m} \alpha_i^D h_i(x^D) \\
&\geq \nabla l_P(x^D)(x^P - x^D) - \sum_{i=1}^{m} \alpha_i^D h_i(x^D) && \text{by (5.15)} \\
&= -\sum_{i=1}^{m} \alpha_i^D \nabla h_i(x^D)(x^P - x^D) - \sum_{i=1}^{m} \alpha_i^D h_i(x^D) && \text{by (5.13)} \\
&\geq -\sum_{i=1}^{m} \alpha_i^D (h_i(x^P) - h_i(x^D)) - \sum_{i=1}^{m} \alpha_i^D h_i(x^D) && \text{by (5.16)} \\
&= -\sum_{i=1}^{m} \alpha_i^D h_i(x^P) \geq 0.
\end{aligned}$$

□

Remark 10.

(a) $l_D(x, \alpha)$ is a lower bound of $l_P(x)$.

(b) For primal feasible v and dual feasible (u, α),

$$\max_{u,\alpha} l_D(u, \alpha) \leq \min_{v} l_p(v).$$

(c) Let $\hat{x}$ be the minimizer of $l_P(x)$ and suppose that $(\hat{x}, \hat{\alpha})$ is dual feasible. Then,

$$l_D(\hat{x}, \hat{\alpha}) = l_P(\hat{x}) + \sum_{i=1}^{m} \hat{\alpha}_i h_i(\hat{x}) \leq l_P(\hat{x}).$$

Moreover, if $l_D(\hat{x}, \hat{\alpha}) = l_P(\hat{x})$, then $\hat{\alpha}_i h_i(\hat{x}) = 0$ for all $i = 1, \ldots, m$. That is,

$$\begin{aligned}
\text{if } \hat{\alpha}_i > 0, \text{ then } h_i(\hat{x}) = 0 \text{ (the constraint } h_i \text{ is active)}, \\
\text{if } h_i(\hat{x}) < 0, \text{ then } \hat{\alpha}_i = 0 \text{ (the constraint } h_i \text{ is inactive)}.
\end{aligned}$$

(d) The conditions,

$$\hat{\alpha}_i h_i(\hat{x}) = 0 \text{ for all } i = 1, \ldots, m, \tag{5.17}$$

are called *complementarity conditions* .

Theorem 14 (Wolfe's strong duality theorem)**.** *Let l_P and h_is be differentiable and convex. Suppose, $\hat{x}$ solves the primal problem. If some constraint qualification conditions hold for the h_i functions then, there exists an $\hat{\alpha} \in \mathbb{R}^m$ such that $(\hat{x}, \hat{\alpha})$ solves the dual problem and $l_P(\hat{x}) = l_D(\hat{x}, \hat{\alpha})$.*