1 Examining the Dual

In the last lecture, we derived a primal standard form convex optimisation problem that can be solved to obtain the threshold function \( h(\vec{x}) = \text{sign}(\vec{w} \cdot \vec{x} + b) \) that maximises the margin of a set of data points \((\vec{x}_1, y_1), \ldots, (\vec{x}_m, y_m)\):

\[
\min_{\vec{w}, b} \frac{1}{2} \| \vec{w} \|^2 \\
\text{s.t.} \quad 1 - y_i (\vec{w} \cdot \vec{x}_i + b) \leq 0, \quad i = 1, \ldots, m.
\]

We also derived the dual of this problem:

\[
\max_{\vec{\alpha}} \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j \\
\text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \ldots, m \\
\sum_{i=1}^{m} \alpha_i y_i = 0.
\]

Since the primal problem is convex (i.e., has a convex objective function and convex inequality constraints), we know that solving the primal is equivalent to solving the dual. Optimal solution \( \vec{w}^*, b^* \), \( \vec{\alpha}^* \) must satisfy the KKT conditions. Let’s see what the KKT conditions tell us about the solutions.

- **Stationarity:**
  Stationarity tells us that the partial derivatives of the Lagrangian with respect to \( \vec{w} \) and \( b \) must be 0 at the optimal solution. As we saw in the last lecture, this implies that

  \[
  \vec{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \vec{x}_i,
  \]

  and

  \[
  \sum_{i=1}^{m} \alpha_i^* y_i = 0.
  \]

  The first equation here tells us how to derive the optimal value \( \vec{w}^* \) of the weights \( \vec{w} \) from the optimal value \( \vec{\alpha}^* \) for \( \vec{\alpha} \) in the dual problem. In particular, the optimal weight vector will be a weighted combination of input points.

  (In order to obtain our classifier, we will need to calculate the optimal value of the offset \( b \) as well. Let’s hold off on the question of how to do this for a moment, and examine what the other KKT conditions tell us about the solution to this problem.)
• **Primal Feasibility:**

Primal feasibility tells us that all primal constraints must be satisfied. In this case, it implies that for all data points \( i \in \{1, \ldots, m\} \),

\[
y_i(\vec{x}_i \cdot \vec{w}^* + b^*) \geq 1,
\]

or

\[
y_i \frac{\vec{x}_i \cdot \vec{w}^* + b^*}{\|\vec{w}^*\|} \geq \frac{1}{\|\vec{w}^*\|}.
\]

In other words, we must have a margin of at least \( 1/\|\vec{w}^*\| \) on the training data.

• **Dual Feasibility:**

Primal feasibility tells us that all dual constraints must be satisfied, or for all \( i \in \{1, \ldots, m\} \), \( \alpha_i^* \geq 0 \). This makes sense. It tells us that our weight vector \( \vec{w}^* = \sum_{i=1}^{m} \alpha_i^* y_i \vec{x}_i \) will only ever put positive weight on positively labelled points, and negative weight on negatively labelled points.

• **Complementary Slackness:**

The implications of this condition are rather interesting. In particular, we get that for all \( i \in \{1, \ldots, m\} \),

\[
\alpha_i^*(y_i(\vec{x}_i \cdot \vec{w}^* + b) - 1) = 0.
\]

Stated another way, for every input point \( \vec{x}_i \), at least one of the corresponding primal constraint and the corresponding dual constraint must be tight, i.e., we must have either \( \alpha_i = 0 \) or \( y_i(\vec{x}_i \cdot \vec{w}^* + b^*) = 1 \).

This implies that the only input points that contribute to the weight vector are those with minimal margin, i.e., those points \( x_i \) such that \( y_i(\vec{x}_i \cdot \vec{w}^* + b^*) = 1 \). These points are referred to as support vectors. The number of support vectors is typically much smaller than the number of data points \( m \), as in Figure 1.

Note that this also gives us a way to solve for the optimal value \( b^* \). We know that for any data point \( x_i \) with \( \alpha_i > 0 \), we must have

\[
b^* = 1/y_i - \vec{x}_i \cdot \vec{w}^* = y_i - \vec{x}_i \cdot \vec{w}^*.
\]

In practice, different points \( x_i \) might give slightly different values of \( b^* \), so it is common to average over all of the support vectors.

2 **Solving The Dual Problem**

Many different techniques can be used to solve the dual problem. We briefly discuss the main idea behind one common hill-climbing technique, Sequential Minimisation Optimisation (SMO). This is explored in more detail in Exercise Sheet 3. For full details on this algorithm, check out John Platt’s book chapter, available for free online. ¹

First, suppose we want to solve an unconstrained optimisation problem of the form \( \max_{\vec{a}} f(\vec{a}) \) for some concave function \( f \). A common technique for doing this is to use a coordinate ascent algorithm. There are different ways to specify the details of coordinate ascent, but the general outline is as follows:

¹[http://research.microsoft.com/apps/pubs/?id=68391](http://research.microsoft.com/apps/pubs/?id=68391)
Figure 1: The support vectors are the input points with minimal margin. Only these points contribute to the weight vector $\vec{w}^*$. 

Initialise $\vec{\alpha}$
Repeat until convergence
Choose some $i$
Set $\alpha_i = \arg\max \hat{f}(\alpha_1, \ldots, \alpha_{i-1}, \alpha_i, \alpha_{i+1}, \ldots, \alpha_m)$

Unfortunately, we can’t apply coordinate ascent directly to the dual problem because of the stationarity condition which tells us that

$$\sum_{i=1}^{m} \alpha_i y_i = 0.$$ 

This condition tells us that there is only one possible value for the $i$th component of $\vec{\alpha}$ given all of the other components, so we can’t optimise components separately.

The SMO algorithm gets around this problem by choosing a pair of components, $\alpha_i$ and $\alpha_j$, at each time step, and optimising them together while maintaining the stationarity constraint. We can easily test for convergence by checking if the KKT conditions are satisfied up to some tolerance parameter.

3 No Perfect Target

So far we have assumed that the data is linearly separable, i.e., that it can be classified perfectly by a linear separator. It turns out that it is easy to modify the optimisation so that it still gives us something reasonable even when there is no perfect target.

3.1 The Soft-Margin Approach

In the soft-margin approach, we introduce a set of “slack variables” $\xi_1, \ldots, \xi_m$, to relax the margin constraint. Our optimisation problem becomes:

$$\begin{align*}
\min_{\vec{w}, b, \vec{\xi}} & \quad \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{m} \xi_i \\
\text{s.t.} & \quad y_i (\vec{w} \cdot \vec{x}_i + b) \geq 1 - \xi_i, \quad \text{for } i = 1, \ldots, m \\
& \quad \xi_i \geq 0, \quad \text{for } i = 1, \ldots, m
\end{align*}$$

where $C$ is a parameter of the algorithm. This allows for occasional failure of the margin condition (i.e., data points for which $y_i (\vec{w} \cdot \vec{x}_i + b) < 1$), but we pay a price in the objective function each time this happens.
It is easy to see that when the optimal solution is found, we have
\[ \xi_i = \max(0, 1 - y_i(\vec{w} \cdot \vec{x}_i + b)) . \]
This is the hinge loss of the \( i \)th data point. We can therefore think of this problem as minimising a weighted combination of the hinge loss \( \sum_{i=1}^{m} \xi_i \) and a regularisation term \( \|\vec{w}\|^2 \) that intuitively penalises complex hypotheses.

To dualise this problem we consider first the Lagrangian. This is defined by
\[
L(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) = \frac{1}{2} \|\vec{w}\|^2 + C \sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i[y_i(\vec{w} \cdot \vec{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{m} \beta_i \xi_i ,
\]
where \( \vec{\alpha} \) and \( \vec{\beta} \) are the Lagrange multipliers.

For fixed values of \( \vec{\alpha} \) and \( \vec{\beta} \) we obtain an expression for \( \min_{\vec{w}, b, \vec{\xi}} L(\vec{w}, b, \vec{\xi}, \vec{\alpha}, \vec{\beta}) \) by setting the gradient of the Lagrangian with respect to the primal variables \( \vec{w}, b, \vec{\xi} \) to zero. This yields
\[
\nabla_{\vec{w}}L = \vec{w} - \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i = 0 \implies \vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i
\]
(2)
\[
\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y_i = 0 \implies \sum_{i=1}^{m} \alpha_i y_i = 0
\]
(3)
\[
\frac{\partial L}{\partial \xi_i} = C - \alpha_i - \beta_i = 0 \implies \alpha_i + \beta_i = C
\]
(4)

Substituting \( \vec{w} = \sum_{i=1}^{m} \alpha_i y_i \vec{x}_i \) into the Lagrangian and using equations (3) and (4) to simplify the resulting expression by eliminating the terms in \( \beta_i \) and \( \xi_i \), we obtain the dual problem:
\[
\max_{\vec{\alpha}} \quad \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} y_i y_j \alpha_i \alpha_j \vec{x}_i \cdot \vec{x}_j
\]
s.t. \( 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, m. \)
\[
\sum_{i=1}^{m} \alpha_i y_i = 0 .
\]

Note that the equation \( \alpha_i + \beta_i = C \) and constraints \( \alpha_i, \beta_i \geq 0 \) were translated to the inequality \( 0 \leq \alpha_i \leq C \).

The objective function that we maximise is the same as in the separable case and we just have the additional constraints that \( \alpha_i \leq C \). Thus to handle data that isn’t linearly separable, we only have to slightly modify our original problem and the same hill climbing techniques work.

This soft max dual optimisation problem is what people typically mean when they refer to the SVM algorithm. This will give good results if the data is “almost” linearly separable.

---

\[ ^2 \]Notice that Equations (3) and (4) are determined by the values of the Lagrange multipliers \( \vec{\alpha} \) and \( \vec{\beta} \), rather than \( \vec{w}, b, \vec{\xi} \). However we can assume without loss of generality that these equations hold since otherwise the value of (1) is unbounded below.