1 Duality

As we have seen in our discussion of kernels, ridge regression can be viewed in two ways: (1) an optimization problem over the weights $\mathbf{w} \in \mathbb{R}^d$ which scales according to the dimensionality of the augmented feature space, and (2) an optimization problem over the weights $\boldsymbol{\alpha} \in \mathbb{R}^n$ which scales according to the the number of training points. These two viewpoints give rise to two equivalent solutions:

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^\top \mathbf{y} \quad \text{and} \quad \mathbf{w}^* = \mathbf{X}^\top (\mathbf{X}\mathbf{X}^\top + \lambda \mathbf{I})^{-1} \mathbf{y}$$

The second (kernelized) expression is much more efficient to calculate when the number of training points n is significantly smaller than the number of augmented features d. Recall that the derivation for the kernelized expression relied on invoking the fundamental theorem of linear algebra and solving for a set of dual variables. While this approach is certainly valid, it may not be applicable for kernelizing all problems. Rather, a more principled approach is to apply Lagrangian duality and solve the dual problem. In this section we will introduce duality for arbitrary optimization problems, and then use duality to derive the kernelized versions for ridge regression and SVMs.

1.1 Primal and Dual Problem

All optimization problems can be expressed in the standard form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f_0(\mathbf{x}) \\ \text{s.t.} \quad & f_i(\mathbf{x}) \leq 0 \quad i = 1, \ldots, m \\ & h_j(\mathbf{x}) = 0 \quad j = 1, \ldots, n \end{aligned} \tag{1}$$

For the purposes of our discussion, assume that $\mathbf{x} \in \mathbb{R}^d$. The components of an optimization problem are:

- The **objective function** $f_0(\mathbf{x})$
- The **inequality constraints**: expressions involving $f_i(\mathbf{x})$
- The **equality constraints**: expressions involving $h_j(\mathbf{x})$

Working with the constraints can be cumbersome and challenging to manipulate, and it would be ideal if we could somehow turn this constrained optimization problem into an unconstrained one. One idea is to re-express the optimization problem into

$$\min_{\mathbf{x}} \mathcal{L}(\mathbf{x})$$

where

$$\mathcal{L}(\mathbf{x}) = \begin{cases} f_0(\mathbf{x}) & \text{if} \quad f_i(\mathbf{x}) \leq 0, \ \forall i \in \{1 \,..\, m\} \quad \text{and} \quad h_j(\mathbf{x}) = 0, \ \forall j \in \{1 \,..\, n\} \\ \infty & \text{otherwise} \end{cases} \tag{2}$$

Note that the unconstrained optimization problem above is equivalent to the original constrained problem. Even though the unconstrained problem considers values that violate the constraints (and therefore are not in the feasible set for the constrained optimization problem), it will effectively ignore them because they are treated as ∞ in a minimization problem.

Even though we are now dealing with an unconstrained problem, it still is difficult to solve the optimization problem, because we still have to deal with all of the casework in the objective function $\mathcal{L}(\mathbf{x})$. In order to solve this issue, we have to introduce dual variables, specifically one set of dual variables for the equality constraints, and one set for the inequality constraints. If we only take into account the dual variables for the equality constraints, the optimization problem now becomes

$$\min_{\mathbf{x}} \max_{\boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\nu})$$

where

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\nu}) = \begin{cases} f_0(\mathbf{x}) + \sum_{j=1}^{n} \nu_j h_j(\mathbf{x}) & \text{if} \quad f_i(\mathbf{x}) \leq 0, \ \forall i \in \{1 \,..\, m\} \\ \text{[illegible]} & \text{otherwise} \end{cases} \tag{3}$$

We are still working with an unconstrained optimization problem, except that now, we are optimizing over two sets of variables: the **primal variables** $\mathbf{x} \in \mathbb{R}^d$ and the **dual variables** $\boldsymbol{\nu} \in \mathbb{R}^n$. Also note that the optimization problem has now become a nested one, with an inner optimization problem the maximizes over the dual variables, and an outer optimization problem that minimizes over the primal variables. Let's examine why this optimization problem is equivalent to the original constrained optimization problem:

- Any $\mathbf{x}$ that violates the inequality constraints is still treated as ∞ by the outer minimization problem over $\mathbf{x}$ and therefore ignored
- For any $\mathbf{x}$ that violates the equality constraints (meaning that $\exists j$ s.t. $h_j(\mathbf{x}) \neq 0$), the inner maximization problem over $\boldsymbol{\nu}$ can choose ν_j as ∞ if $h_j(\mathbf{x}) > 0$ (or ν_j as $-\infty$ if $h_j(\mathbf{x}) < 0$) to cause the inner maximization to go to ∞, therefore being ignored by the outer minimization over $\mathbf{x}$
- For any $\mathbf{x}$ that does not violate any of the equality or inequality constraints, the inner maximization problem over $\boldsymbol{\nu}$ is simply equal to $f_0(\mathbf{x})$

This solution comes at a cost — in an effort to remove the equality constraints, we had to add in dual variables, one for each equality constraint. With this in mind, let's try to do the same for the inequality constraints. Adding in dual variable λ_i to represent each inequality constraint, we now have

$$\begin{aligned} \min_{\mathbf{x}} \max_{\boldsymbol{\lambda}, \boldsymbol{\nu}} \quad & \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j h_j(\mathbf{x}) \\ \text{s.t.} \quad & \lambda_i \geq 0 \quad i = 1, \ldots, m \end{aligned} \tag{4}$$

For convenience, we can place the constraints involving $\boldsymbol{\lambda}$ into the optimization variable.

$$\min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j h_j(\mathbf{x}) \tag{5}$$

This optimization problem above is otherwise known as the **primal** (not to be confused with the *primal variables*), and its optimal value is indeed *equivalent* to that of the original constrained optimization problem.

$$p^* = \min_{\mathbf{x}} \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{6}$$

We can verify that this is indeed the case:

- For any $\mathbf{x}$ that violates the inequality constraints (meaning that $\exists i \in \{1 \,..\, m\}$ s.t. $f_i(\mathbf{x}) > 0$), the inner maximization problem over $\boldsymbol{\lambda}$ can choose λ_i as ∞ to cause the inner maximization go to ∞, therefore being ignored by the outer minimization over $\mathbf{x}$
- For any $\mathbf{x}$ that violates the equality constraints (meaning that $\exists j$ s.t. $h_j(\mathbf{x}) \neq 0$), the inner maximization problem over $\boldsymbol{\nu}$ can choose ν_j as ∞ if $h_j(\mathbf{x}) > 0$ (or ν_j as $-\infty$ if $h_j(\mathbf{x}) < 0$) to cause the inner maximization go to ∞, therefore being ignored by the outer minimization over $\mathbf{x}$
- For any $\mathbf{x}$ that does not violate any of the equality or inequality constraints, in the inner maximization problem over $\boldsymbol{\nu}$, the expression $\sum_{j=1}^{n} \nu_j h_j(\mathbf{x})$ evaluates to 0 no matter what the value of $\boldsymbol{\nu}$ is, and in the inner maximization problem over $\boldsymbol{\lambda}$, the expression $\sum_{i=1}^{m} \lambda_i f_i(\mathbf{x})$ can at maximum be 0, because λ_i is constrained to be non-negative, and $f_i(\mathbf{x})$ is non-positive. Therefore, at best, the maximization problem sets $\lambda_i f_i(\mathbf{x}) = 0$, and

$$\max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x})$$

In its full form, the objective $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is called the **Lagrangian**, and it takes into account the unconstrained set of primal variables $\mathbf{x} \in \mathbb{R}^d$, the constrained set of dual variables $\boldsymbol{\lambda} \in \mathbb{R}^n$ corresponding to the inequality constraints, and the unconstrained set of dual variables $\boldsymbol{\nu} \in \mathbb{R}^m$ corresponding to the equality constraints. Note that our dual variables λ_i are in fact constrained, so ultimately we were not able to turn the original optimization problem into an unconstrained one, but our constraints are much simpler than before.

The **dual** of this optimization problem is still over the same optimization objective, except that now we swap the order of the maximization of the dual variables and the minimization of the primal variables.

$$d^* = \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{7}$$

The dual is effectively a maximization problem (over the dual variables):

$$d^* = \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{8}$$

where

$$g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \tag{9}$$

The dual is very useful to work with, because now the inner optimization problem over $\mathbf{x}$ is an unconstrained problem! Furthermore, the dual $g(\boldsymbol{\lambda}, \boldsymbol{\nu})$ is always a concave function, regardless of the primal objective function or its constraints. This is because the dual is a pointwise minimum of concave functions, which itself is a concave function. Specifically $g(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is a pointwise minimum of functions $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ that are affine in the dual variables (which are both concave and convex at the same time).

1.2 Strong Duality and KKT Conditions

Let's examine the relationship between the primal and dual problem. It is *always* true that the solution to the primal problem is at least as large as the solution to the dual problem:

$$p^* \geq d^* \tag{10}$$

This condition is know as **weak duality**.

Proof. We know that

$$\forall \mathbf{x}, \boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu} \quad \max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \mathcal{L}(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \geq \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \geq \min_{\tilde{\mathbf{x}}} \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

More compactly,

$$\forall \mathbf{x}, \boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu} \quad \max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \mathcal{L}(\mathbf{x}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \geq \min_{\tilde{\mathbf{x}}} \mathcal{L}(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu})$$

Since this is true for all $\mathbf{x}, \boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}$ this is true in particular when we set

$$\mathbf{x} = \arg\min_{\tilde{\mathbf{x}}} \max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$$

and

$$\boldsymbol{\lambda}, \boldsymbol{\nu} = \arg\max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \min_{\tilde{\mathbf{x}}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$$

We therefore know that

$$p^* = \min_{\tilde{\mathbf{x}}} \max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \geq \max_{\tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}} \min_{\tilde{\mathbf{x}}} \mathcal{L}(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = d^*$$

□

The difference $p^* - d^*$ is known as the **duality gap**. In the case of **strong duality**, the duality gap is 0. That is, we can swap the order of the minimization and maximization and up with the same optimal value:

$$p^* = d^* \tag{11}$$

There are several useful theorems detailing the existence of strong duality, such as **Slater's theorem**, which states that if the primal problem is convex, and there exists an $\mathbf{x}$ that can *strictly* meet the inequality constraints and meet the equality constraints, then strong duality holds. Given that strong duality holds, the **Karush-Kuhn-Tucker (KKT) conditions** can help us find the solution to the dual variables of the optimization problem. The KKT conditions are composed of:

1. Primal feasibility (inequalities)

$$f_i(\mathbf{x}) \leq 0, \ \forall i \in \{1 \mathinner{..} m\}$$

2. Primal feasibility (equalities)

$$h_j(\mathbf{x}) = 0, \ \forall j \in \{1 \mathinner{..} n\}$$

3. Dual feasibility

$$\lambda_i \geq 0, \ \forall i \in \{1 \mathinner{..} m\}$$

4. Complementary Slackness

$$\lambda_i f_i(\mathbf{x}) = 0, \ \forall i \in \{1 \mathinner{..} m\}$$

5. Stationarity

$$\nabla_{\mathbf{x}} f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} f_i(\mathbf{x}) + \sum_{j=1}^{n} \nu_j \nabla_{\mathbf{x}} h_j(\mathbf{x}) = 0$$

Let's see how the KKT conditions relate to strong duality.

Theorem 1. *If $\mathbf{x}^*$ and $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ are the primal and dual solutions respectively, with zero duality gap (i.e. strong duality holds), then $\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ also satisfy the KKT conditions.*

Proof. KKT conditions 1, 2, 3 are trivially true, because the primal solution $\mathbf{x}^*$ must satisfy the primal constraints, and the dual solution $\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*$ must satisfy the dual constraints. Now, let's prove conditions 4 and 5. We know that since strong duality holds, we can say that

$$\begin{aligned}
p^* = f_0(\mathbf{x}^*) &= g(\boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = d^* && (12)\\
&= \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) && (13)\\
&\leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) && (14)\\
&= f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) + \cancel{\sum_{j=1}^{n} \nu_j^* h_j(\mathbf{x}^*)} && (15)\\
&= f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) && (16)\\
&\leq f_0(\mathbf{x}^*) && (17)
\end{aligned}$$

We cancel the terms involving $h_j(\mathbf{x}^*)$ because we know that the primal solution must satisfy $h_j(\mathbf{x}^*) = 0$. Furthermore, we know that $\lambda_i^* f_i(\mathbf{x}^*) \leq 0$, because $\lambda_i^* \geq 0$ in order to satisfy the dual constraints, and $f_i(\mathbf{x}^*) \leq 0$ in order to satisfy the primal constraints. Since we established that $f_0(\mathbf{x}^*) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq \mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) \leq f_0(\mathbf{x}^*)$, we know that all of the inequalities

hold with equality and therefore $\mathcal{L}(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}^*, \boldsymbol{\nu}^*)$. This implies KKT condition 5 (stationarity), that

$$\nabla_{\mathbf{x}} f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_{\mathbf{x}} f_i(\mathbf{x}^*) + \sum_{j=1}^{n} \nu_j^* \nabla_{\mathbf{x}}^* h_j(\mathbf{x}^*) = 0$$

Finally, note that due to the equality $f_0(\mathbf{x}^*) + \sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) = f_0(\mathbf{x}^*)$, we know that $\sum_{i=1}^m \lambda_i^* f_i(\mathbf{x}^*) = 0$. This combined with the fact that $\forall i \quad \lambda_i^* f_i(\mathbf{x}^*) \leq 0$, establishes KKT condition 4 (complementary slackness):

$$\lambda_i^* f_i(\mathbf{x}^*) = 0, \ \forall i \in \{1 \,..\, m\}$$

□

The theorem above establishes that in the presence of strong duality, if the solutions are optimal, then they satisfy the KKT conditions. Let's prove a statement that is almost (but not quite) the converse, which will be much more helpful for solving optimization problems.

Theorem 2. *If $\bar{\mathbf{x}}$ and $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}$ satisfy the KKT conditions, and the primal problem is convex, then they are the optimal solutions to the primal and dual problems with zero duality gap.*

Proof. If $\bar{\mathbf{x}}$ and $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}$ satisfy KKT conditions 1, 2, 3 we know that they are at least feasible for the primal and dual problem. From the KKT stationarity condition we know that

$$\nabla_{\mathbf{x}} f_0(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\lambda}_i \nabla_{\mathbf{x}} f_i(\bar{\mathbf{x}}) + \sum_{j=1}^{n} \bar{\nu}_j \nabla_{\mathbf{x}} h_j(\bar{\mathbf{x}}) = 0$$

Since the primal problem is convex, we know that $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu})$ is convex in $\mathbf{x}$, and if the gradient of $\mathcal{L}(\mathbf{x}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}})$ at $\bar{\mathbf{x}}$ is $\mathbf{0}$, we know that

$$\bar{\mathbf{x}} = \arg\min_{\mathbf{x}} \ \mathcal{L}(\mathbf{x}, \bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}})$$

Therefore, we know that the optimal primal values for the primal problem optimize the inner optimization problem of the dual problem, and

$$g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}) = f_0(\bar{\mathbf{x}}) + \sum_{i=1}^{m} \bar{\lambda}_i f_i(\bar{\mathbf{x}}) + \sum_{j=1}^{n} \bar{\nu_j} h_j(\bar{\mathbf{x}})$$

By the primal feasibility conditions for $h_j(\mathbf{x})$ and the complementary slackness condition, we know that

$$g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}) = f_0(\bar{\mathbf{x}})$$

Now, all we have to do is to prove that $\bar{\mathbf{x}}$ and $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}$ are primal and dual optimal, respectively. Note that since weak duality always holds, we know that

$$p^* \geq d^* = \max_{\boldsymbol{\lambda} \geq \mathbf{0}, \boldsymbol{\nu}} g(\boldsymbol{\lambda}, \boldsymbol{\nu}) \geq g(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}), \quad \forall \tilde{\boldsymbol{\lambda}} \geq \mathbf{0}, \tilde{\boldsymbol{\nu}}$$

Since we know that $p^* \geq g(\boldsymbol{\lambda}, \boldsymbol{\nu})$, we can also say that

$$f_0(\mathbf{x}) - p^* \leq f_0(\mathbf{x}) - g(\boldsymbol{\lambda}, \boldsymbol{\nu})$$

And if we have that $f_0(\bar{\mathbf{x}}) = g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}})$ as we deduced earlier, then

$$f_0(\bar{\mathbf{x}}) - p^* \leq f_0(\bar{\mathbf{x}}) - g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}) = 0 \implies p^* \geq f_0(\bar{\mathbf{x}})$$

Since p^* is the minimum value for the primal problem, we can go further by saying that $p^* \geq f_0(\bar{\mathbf{x}})$ holds with equality and

$$p^* = f_0(\bar{\mathbf{x}}) = g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}) \leq d^*$$

since it always holds that $p^* \geq d^*$ we conclude that

$$p^* = f_0(\bar{\mathbf{x}}) = g(\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}) = d^*$$

Therefore, we have proven that $\bar{\mathbf{x}}$ and $\bar{\boldsymbol{\lambda}}, \bar{\boldsymbol{\nu}}$ are primal and dual optimal respectively, with zero duality gap. We eventually arrived at the conclusion that strong duality does indeed hold. □

Let's pause for a second to understand what we've found so far. Given an optimization problem, its primal problem is an optimization problem over the primal variables, and its dual problem is an optimization problem over the dual variables. If strong duality holds, then we can solve the dual problem and arrive at the same optimal value. In order to solve the dual, we have to first solve the unconstrained inner optimization problem over the primal variables and then solve the constrained outer optimization problem over the dual variables. But how do we even know in the first place that strong duality holds? This is where KKT comes into play. If the the primal problem is convex and the KKT conditions hold, we can solve for the dual variables easily and also verify strong duality does indeed hold. We shall do just that, in our discussion of dual ridge regression and dual SVMs.

1.3 Dual Ridge Regression

Let's derive kernel ridge regression again, using duality this time. Recall the unconstrained ridge regression formulation:

$$\min_{\mathbf{w}} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2 + \lambda \|\mathbf{w}\|^2$$

This formulation is not conducive to dualization, because it lacks constraints. We will add constraints by introducing a dummy variable $\mathbf{z} = \mathbf{X}\mathbf{w} - \mathbf{y}$ that corresponds to equality constraints:

$$\begin{aligned} \min_{\mathbf{w},\mathbf{z}} \quad & \|\mathbf{z}\|^2 + \lambda \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & \mathbf{z} = \mathbf{X}\mathbf{w} - \mathbf{y} \end{aligned} \tag{18}$$

Now we proceed to forming the dual problem. For the purposes of notation, note that we are using $\boldsymbol{\alpha}$ in place of $\boldsymbol{\nu}$, and there are no dual variables corresponding to $\boldsymbol{\lambda}$ because there are no inequality constraints. The Lagrangian is

$$\mathcal{L}(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha}) = \|\mathbf{z}\|^2 + \lambda \|\mathbf{w}\|^2 + \boldsymbol{\alpha}^\top (\mathbf{X}\mathbf{w} - \mathbf{y} - \mathbf{z})$$

The dual problem is

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) \tag{19}$$

where

$$g(\boldsymbol{\alpha}) = \min_{\mathbf{w},\mathbf{z}} \|\mathbf{z}\|^2 + \lambda\|\mathbf{w}\|^2 + \boldsymbol{\alpha}^\top(\mathbf{X}\mathbf{w} - \mathbf{y} - \mathbf{z}) \tag{20}$$

Since the $g(\boldsymbol{\alpha})$ is a convex minimization problem over the variables $\mathbf{w}$ and $\mathbf{z}$, we can simply set the derivative to $\mathbf{0}$ w.r.t. $\mathbf{w}$ and $\mathbf{z}$:

- $\nabla_{\mathbf{w}}\mathcal{L} = 2\lambda\mathbf{w} + \mathbf{X}^\top\boldsymbol{\alpha} = \mathbf{0} \implies \mathbf{w}^*(\boldsymbol{\alpha}) = -\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha}$. This tells us that $\mathbf{w}^*$ is going to be a linear combination of the $\mathbf{x}_i$'s.
- $\nabla_{\mathbf{z}}\mathcal{L} = 2\mathbf{z} - \boldsymbol{\alpha} = \mathbf{0} \implies \mathbf{z}^*(\boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{\alpha}$.

Plugging these optimal values back into the optimization problem, we have that

$$g(\boldsymbol{\alpha}) = \min_{\mathbf{w},\mathbf{z}} \mathcal{L}(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha}) \tag{21}$$

$$= \mathcal{L}(\mathbf{w}^*(\boldsymbol{\alpha}), \mathbf{z}^*(\boldsymbol{\alpha}), \boldsymbol{\alpha}) \tag{22}$$

$$= \left\| -\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha} \right\|^2 + \lambda\left\| -\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha} \right\|^2 + \left(-\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha}\right)^\top\left(\mathbf{X}\left(-\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha}\right) - \mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}\right) \tag{23}$$

$$= -\frac{1}{4}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{1}{4\lambda}\boldsymbol{\alpha}^\top\mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\mathbf{y} \tag{24}$$

Now, the dual problem is

$$\max_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} -\frac{1}{4}\boldsymbol{\alpha}^\top\boldsymbol{\alpha} - \frac{1}{4\lambda}\boldsymbol{\alpha}^\top\mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha} - \boldsymbol{\alpha}^\top\mathbf{y}$$

Note that this problem is a maximization over a concave problem (similar to a minimization over a convex problem) and we can take the derivative w.r.t $\boldsymbol{\alpha}$ and set it to $\mathbf{0}$:

$$\nabla_{\boldsymbol{\alpha}} g(\boldsymbol{\alpha}) = -\frac{1}{2}\boldsymbol{\alpha} - \frac{1}{2\lambda}\mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha} - \mathbf{y} = \mathbf{0} \implies \boldsymbol{\alpha}^* = -2\lambda(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}$$

The optimal $\mathbf{w}^*$ is therefore given by

$$\mathbf{w}^* = -\frac{1}{2\lambda}\mathbf{X}^\top\boldsymbol{\alpha}^* = \mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y}$$

Which exactly matches the expression we previously derived for kernel ridge regression! Note that while this solution is dual optimal, it may not be optimal for the primal problem. In order to ensure that it is primal optimal, we need to establish that strong duality holds. In this case the primal problem is convex, so we simply need to ensure that the KKT conditions hold. Since we are not dealing with any inequality conditions here, the only applicable conditions are primal feasibility for the equalities and stationarity. Indeed the primal equality constraints are met, since

$$\mathbf{X}\mathbf{w}^* - \mathbf{y} - \mathbf{z}^* = -\frac{1}{2\lambda}\mathbf{X}\mathbf{X}^\top\boldsymbol{\alpha}^* - \mathbf{y} - \frac{1}{2}\boldsymbol{\alpha}^*$$

$$
\begin{aligned}
&= -\frac{1}{2\lambda}(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})\boldsymbol{\alpha}^* - \mathbf{y} \\
&= (\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})(\mathbf{X}\mathbf{X}^\top + \lambda\mathbf{I})^{-1}\mathbf{y} - \mathbf{y} \\
&= \mathbf{0}
\end{aligned}
$$

We already showed the stationarity conditions are met, when we were solving $g(\boldsymbol{\alpha}) = \min_{\mathbf{w},\mathbf{z}} \mathcal{L}(\mathbf{w}, \mathbf{z}, \boldsymbol{\alpha})$. We conclude that $\mathbf{w}^*$ is indeed the optimal solution to the primal problem.

1.4 Dual SVMs

Previously in our investigation of SVMs, we formulated a constrained optimization problem that we can solve to find the optimal parameters for our hyperplane decision boundary. Recall the setup of soft-margin SVMs:

- y_i's: ± 1, representing positive or negative class
- $\mathbf{x}_i$'s: feature vectors in $\mathbb{R}^d$

- ξ_i's: slack variables representing how much an $\mathbf{x}_i$ is allowed to violate the margin
- C: a hyperparameter describing how severely we penalize slack

- The optimization problem for $\mathbf{w} \in \mathbb{R}^d$ and $b \in \mathbb{R}$, the parameters of the SVM:

$$
\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & y_i(\mathbf{w}^\top\mathbf{x}_i - b) \geq 1 - \xi_i \quad \forall i \\
& \xi_i \geq 0 \quad \forall i
\end{aligned}
\tag{25}
$$

Now, let's investigate the dual of this problem. The primal problem in standard form is

$$
\begin{aligned}
\min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\
\text{s.t.} \quad & (1 - \xi_i) - y_i(\mathbf{w}^\top\mathbf{x}_i - b) \leq 0 \quad \forall i \\
& -\xi_i \leq 0 \quad \forall i
\end{aligned}
\tag{26}
$$

Let's identify the primal and dual variables for the SVM problem. We will have

- Primal variables $\mathbf{w}$, b, and ξ_i
- Dual variables α_i corresponding to each constraint of the form $y_i(\mathbf{w}^\top\mathbf{x}_i - b) \geq 1 - \xi_i$
- Dual variables β_i corresponding to each constraint of the form $\xi_i \geq 0$

For the purposes of notation, note that we are using $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in place of $\boldsymbol{\lambda}$, and there are no dual variables corresponding to $\boldsymbol{\nu}$ because there are no equality constraints. The Lagrangian for the SVM problem is

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^n \xi_i + \sum_{i=1}^n \alpha_i((1-\xi_i) - y_i(\mathbf{w}^\top \mathbf{x}_i - b)) + \sum_{i=1}^n \beta_i(-\xi_i) \tag{27}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i(\mathbf{w}^\top \mathbf{x}_i - b) + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n (C - \alpha_i - \beta_i)\xi_i \tag{28}$$

Thus, the dual is

$$\max_{\boldsymbol{\alpha}\geq 0, \boldsymbol{\beta}\geq 0} g(\boldsymbol{\alpha}, \boldsymbol{\beta}) \tag{29}$$

where

$$g(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \min_{\mathbf{w}, b, \boldsymbol{\xi}} \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i y_i(\mathbf{w}^\top \mathbf{x}_i - b) + \sum_{i=1}^n \alpha_i + \sum_{i=1}^n (C - \alpha_i - \beta_i)\xi_i \tag{30}$$

Let's use the KKT conditions to find the optimal dual variables. Verify that the primal problem is convex in the primal variables. We know that from the stationarity conditions, evaluated at the optimal dual values α^* and β^*, and the optimal primal values $\mathbf{w}^*, b^*, \xi_i^*$:

$$\frac{\partial \mathcal{L}}{\partial w_i} = \frac{\partial \mathcal{L}}{\partial b} = \frac{\partial \mathcal{L}}{\partial \xi_i} = 0$$

- $\nabla_{\mathbf{w}}\mathcal{L} = \mathbf{w}^* - \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i = \mathbf{0} \implies \mathbf{w}^* = \sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i$. This tells us that $\mathbf{w}^*$ is going to be a weighted combination of the positive-class $\mathbf{x}_i$'s and negative-class $\mathbf{x}_i$'s.
- $\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^n \alpha_i^* y_i = 0$. This tells us that the weights α_i^* will be equally distributed among positive- and negative- class training points.
- $\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i^* - \beta_i^* = 0 \implies 0 \leq \alpha_i^* \leq C$. This tells us that the weights α_i^* are restricted to being less than the hyperparameter C.

Verify that the other KKT also hold, establishing strong duality. Using these observations, we can eliminate some terms of the dual problem.

$$\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i^* y_i(\mathbf{w}^\top \mathbf{x}_i - b) + \sum_{i=1}^n \alpha_i^* + \sum_{i=1}^n (C - \alpha_i^* - \beta_i^*)\xi_i \tag{31}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i^* y_i(\mathbf{w}^\top \mathbf{x}_i) + b\underbrace{\sum_{i=1}^n \alpha_i^* y_i}_{=0} + \sum_{i=1}^n \alpha_i^* + \underbrace{\sum_{i=1}^n (C - \alpha_i^* - \beta_i^*)\xi_i}_{=0} \tag{32}$$

$$= \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^n \alpha_i^* y_i(\mathbf{w}^\top \mathbf{x}_i) + \sum_{i=1}^n \alpha_i^* \tag{33}$$

Since the primal problem is convex, from the KKT conditions we have that the optimal primal variables $\mathbf{w}^*, b^*, \boldsymbol{\xi}^*$ minimize $\mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$:

$$g(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \min_{\mathbf{w},b,\boldsymbol{\xi}} \mathcal{L}(\mathbf{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \tag{34}$$

$$= \mathcal{L}(\mathbf{w}^*, b^*, \boldsymbol{\xi}^*, \boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \tag{35}$$

$$= \frac{1}{2}\|\sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i\|^2 - \sum_{i=1}^n \alpha_i^* y_i ((\sum_{j=1}^n \alpha_j^* y_j \mathbf{x}_j)^\top \mathbf{x}_i) + \sum_{i=1}^n \alpha_i^* \tag{36}$$

$$= \frac{1}{2}\|\sum_{i=1}^n \alpha_i^* y_i \mathbf{x}_i\|^2 - \sum_{i=1}^n (\alpha_i^* y_i \mathbf{x}_i^\top (\sum_{j=1}^n \alpha_j^* y_j \mathbf{x}_j)) + \sum_{i=1}^n \alpha_i^* \tag{37}$$

$$= \mathbf{1}^\top \boldsymbol{\alpha}^* - \frac{1}{2}\boldsymbol{\alpha}^{*\top} \mathbf{Q} \boldsymbol{\alpha}^* \tag{38}$$

where $Q_{ij} = y_i(\mathbf{x}_i^\top \mathbf{x}_j)y_j$ (and $\mathbf{Q} = (\text{diag } \mathbf{y})\mathbf{X}\mathbf{X}^\top(\text{diag } \mathbf{y})$).

Now, we can write the final form of the dual, which is only in terms of $\boldsymbol{\alpha}$ and $\mathbf{X}$ and $\mathbf{y}$ (Note that we have eliminated all references to $\boldsymbol{\beta}$):

$$\begin{aligned} \text{maximize}_{\boldsymbol{\alpha}} \quad & \mathbf{1}^\top \boldsymbol{\alpha} - \frac{1}{2}\boldsymbol{\alpha}^\top \mathbf{Q} \boldsymbol{\alpha} \\ \text{s.t.} \quad & \sum_{i=1}^n \alpha_i y_i = 0 \\ & 0 \leq \alpha_i \leq C, \quad i = 1, \ldots, n \end{aligned} \tag{39}$$

Remember to account for the constraints $\sum_{i=1}^n \alpha_i y_i = 0$ and $0 \leq \alpha_i \leq C$ that arise from the stationarity conditions. After all of this effort, we have managed to turn a minimization problem over the primal variables into a maximization problem over the dual variables. One might ask, why go through the effort to formulate and solve the dual problem instead? For one, the dual is an optimization problem over the number of training points n rather than the number of augmented features d, making it particularity attractive when $n \ll d$. Second, it incorporates the term $\mathbf{X}\mathbf{X}^\top$ which is simply the Gram matrix $\mathbf{K}$ of kernel evaluations among all pairs of training points. We can apply the kernel trick to form this Gram matrix, effectively relying on the the dimensionality of the raw feature space rather than the augmented feature space. These are more or less the exact same justifications for kernel ridge regression.

1.4.1 Geometric intuition

We've formulated the dual SVM problem and used the KKT conditions to formulate an equivalent optimization problem, but what do these dual values α_i even mean? That's a good question!

We know that given optimal primal and dual values, the following KKT conditions are enforced:

- Stationarity

$$C - \alpha_i^* - \beta_i^* = 0$$

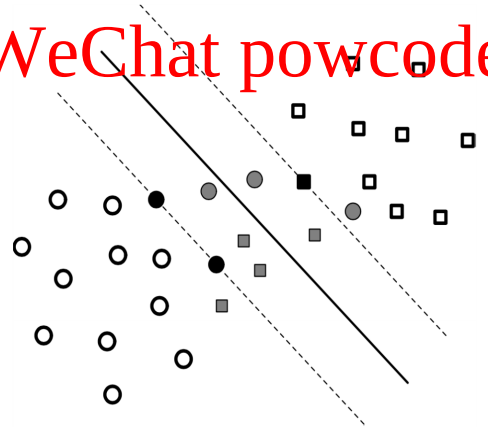
- Complementary slackness

$$\alpha_i^* \cdot ((1 - \xi_i^*) - y_i(\mathbf{w}^{*\top}\mathbf{x}_i - b^*)) = 0$$

$$\beta_i^* \cdot \xi_i^* = 0$$

Here are some noteworthy relationships between α_i and the properties of the SVMs:

- Case 1: $\alpha_i^* = 0$. In this case, we know $\beta_i^* = C$, which is nonzero, and therefore $\xi_i^* = 0$. That is, if for point i we have that $\alpha_i^* = 0$ by the dual problem, then we know that there is no slack given to this point. Looking at the other complementary slackness condition, this makes sense because if $\alpha_i^* = 0$, then $y_i(\mathbf{w}^{*\top}\mathbf{x}_i - b^*) - (1 - \xi_i^*)$ may be any value, and if we're minimizing the sum of our ξ_is, we should have $\xi_i^* = 0$. So, point i lies **on or outside the margin.**
- Case 2: α_i^* is nonzero. If this is the case, then we know $\beta_i^* = C - \alpha_i^* \geq 0$
  - Case 2.1: $\alpha_i^* = C$. If this is the case, then we know $\beta_i^* = 0$, and therefore ξ_i^* may be exactly 0 or nonzero. So, point i lies **on or violates the margin.**
  - Case 2.2: $0 < \alpha_i^* < C$. In this case, then ξ_i^* is nonzero and $\xi_i^* = 0$. But this is different from Case 1 because with α_i^* nonzero, we can divide by α_i^* in the complementary slackness condition and arrive at the fact that $1 - y_i(\mathbf{w}^{*\top}\mathbf{x}_i - b^*) = 0 \implies y_i(\mathbf{w}^{*\top}\mathbf{x}_i - b^*) = 1$, which means $\mathbf{x}_i$ lies exactly on the margin determined by $\mathbf{w}^*$ and b^*. So, point i lies **on the margin.**



$\alpha_i = 0 \Rightarrow$ $y_i f(x_i) \geq 1$: on or outside the margin
$0 < \alpha_i < C \Rightarrow$ $y_i f(x_i) = 1$: on the margin
$\alpha_i = C \Rightarrow$ $y_i f(x_i) \leq 1$: on or inside the margin
$\alpha_i = 0 \Leftarrow$ $y_i f(x_i) > 1$: outside the margin
$\alpha_i = C \Leftarrow$ $y_i f(x_i) < 1$: inside the margin

Using this information, let's reconstruct the optimal primal values $\mathbf{w}^*, b^*, \xi_i^*$ from the optimal dual

values $\boldsymbol{\alpha}^*$:

$$\begin{aligned}
\mathbf{w}^* &= \sum_{i=1}^{n} \alpha_i^* y_i \mathbf{x}_i \\
b^* &= \mathbf{w}^{*\top} \mathbf{x}_i - y_i \quad \text{if } 0 < \alpha_i^* < C \\
\xi_i^* &= \begin{cases} 1 - y_i(\mathbf{w}^{*\top} \mathbf{x}_i - b^*) \text{ if } \alpha_i^* = C, \\ 0 \text{ otherwise} \end{cases}
\end{aligned} \tag{40}$$

The principal takeaway is that the optimal $\mathbf{w}^*$ is a linear combination of the training points for which the corresponding dual weight α_i is non-zero. Such points are called **support vectors**, because they determine the optimal $\mathbf{w}^*$. There is a special relationship between the values of α_i and the position of $\mathbf{x}_i$ relative to the margin. All training points that violate the decision boundary have $\alpha_i > 0$ and are thus support vectors, while all training points that strictly do not violate the decision boundary (meaning that they do not lie on the boundary) have $\alpha_i = 0$ and are not support vectors. For training points which lie exactly on the boundary, some may have $\alpha_i > 0$ and some may have $\alpha_i = 0$; only the points that are critical to determining the decision boundary have $\alpha_i > 0$ and are thus support vectors. Intuitively, there are very few support vectors compared to the total number of training points, meaning that the dual vector $\boldsymbol{\alpha}^*$ is *sparse*. This is advantageous when predicting class for a test point:

$$\mathbf{w}^{*\top} \phi(\mathbf{x}) + b^* = \sum_{i=1}^{n} \alpha_i^* y_i \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}) + b^* = \sum_{i=1}^{n} \alpha_i^* y_i k(\mathbf{x}_i, \mathbf{x}) + b^*$$

We only have to make $m \ll n$ kernel evaluations to predict a test point, where m is the number of support vectors. It should now be clear why the dual SVM problem is so useful: it allows us to use the kernel trick to eliminate dependence on the dimensionality of the argument feature space, while also allowing us to discard most training points because they have dual weight 0.