Understanding Duality and Lagrangians in Optimization

Rishabh Iyer

March 19, 2024

1 Introduction

Duality and Lagrangians play a crucial role in optimization, offering insights into the properties of optimization problems and providing methods for finding solutions. This article explores the formulation of the primal problem, the construction of the Lagrangian, the derivation of the dual problem, and the relationship between primal and dual solutions.

2 Dualiy and Lagrangian

2.1 The Primal Problem

Consider a general optimization problem (the primal problem) defined as:

$$
\begin{aligned}
\text{minimize} \quad & f_0(x) \\
\text{subject to} \quad & f_i(x) \leq 0, \quad i = 1, \ldots, M \\
& h_i(x) = 0, \quad i = 1, \ldots, P
\end{aligned}
$$

where $x \in \mathbb{R}^n$ is the optimization variable, $f_0(x)$ is the objective function, $f_i(x) \leq 0$ are M inequality constraints, and $h_i(x) = 0$ are P equality constraints.

2.2 Forming the Lagrangian

The Lagrangian $L(x, \lambda, \nu)$ for the primal problem is formed by combining the objective function with the constraints, weighted by Lagrange multipliers λ_i for inequality constraints and ν_i for equality constraints:

$$
L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{M} \lambda_i f_i(x) + \sum_{i=1}^{P} \nu_i h_i(x)
$$

2.3   The Dual Problem

The dual problem is derived from the Lagrangian by minimizing $L(x, \lambda, \nu)$ with respect to x and then maximizing with respect to the Lagrange multipliers λ and ν, under the condition that $\lambda_i \geq 0$. The dual function is defined as:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu)$$

The dual problem is then:

$$\begin{aligned} \text{maximize} \quad & g(\lambda, \nu) \\ \text{subject to} \quad & \lambda_i \geq 0, \quad i = 1, \ldots, M \end{aligned}$$

3   Example

Primal Problem

Consider a simple linear programming problem aimed at minimizing a linear objective function subject to linear inequality constraints. The primal problem is formulated as follows:

Minimize the objective function:

$$f(x) = 2x_1^2 + 3x_2^2$$

Subject to the constraints:

$$x_1 + 2x_2 \geq 4,$$
$$3x_1 + x_2 \geq 3$$

Forming the Lagrangian

To form the Lagrangian of this problem, we introduce Lagrange multipliers λ_1 and λ_2 for the inequality constraints. The Lagrangian $L(x, \lambda)$ is given by:

$$L(x, \lambda) = 2x_1^2 + 3x_2^2 + \lambda_1(4 - x_1 - 2x_2) + \lambda_2(3 - 3x_1 - x_2) \tag{1}$$

with $\lambda_1, \lambda_2 \geq 0$.

The Dual Problem

The dual problem is derived by maximizing the Lagrangian with respect to the Lagrange multipliers under the condition that the multipliers are non-negative. The dual function $g(\lambda)$ is obtained by minimizing $L(x, \lambda)$ with respect to x, leading to:

$$g(\lambda) = \inf_x L(x, \lambda).$$

We can take the partial derivatives of L with respect to x_1 and x_2 and set it to zero. We can solve this and get the value of x_1, x_2 as a function of λ. Specifically:

$$4x_1 - \lambda_1 - 3\lambda_2 = 0$$
$$6x_2 - 2\lambda_1 - \lambda_2 = 0$$

This means $x_1 = (\lambda_1 + 3\lambda_2)/4$ and $x_2 = (2\lambda_1 + \lambda_2)/6$. We substitute x_1, x_2 back into the Lagrangian and we get $g(\lambda)$.

4 Weak and Strong Duality

Duality theory plays a central role in optimization, offering profound insights into the relationship between a primal optimization problem and its dual. Understanding the distinction between weak and strong duality is crucial for grasping the efficiency and applicability of dual problem solutions in various contexts.

4.1 Weak Duality

Weak duality refers to the property that the optimal value of the dual problem is always less than or equal to the optimal value of the primal problem for minimization problems (or greater than or equal for maximization problems). Formally, if p^* is the optimal value of the primal problem and d^* is the optimal value of the dual problem, then $d^* \leq p^*$ for minimization problems. This property holds for all optimization problems, regardless of whether they are convex.

Weak duality provides a mechanism to derive lower bounds on the optimal value of the primal problem by solving the dual problem, which can be significantly simpler or more computationally efficient in certain cases.

4.2 Strong Duality

Strong duality goes a step further by stating that the optimal values of the primal and dual problems are equal, $d^* = p^*$, under certain conditions. This equality implies that solving the dual problem is not only a matter of finding a lower bound but actually yields the exact optimal value of the primal problem.

Strong duality typically holds for convex optimization problems, provided they satisfy a condition known as Slater's condition. Slater's condition requires that there exists a feasible point for the primal problem that strictly satisfies all inequality constraints (i.e., the constraints are satisfied with strict inequalities). For linear programming problems, strong duality always holds if the primal problem is feasible and bounded.

4.3 Implications of Strong Duality

The presence of strong duality has several important implications:

- It validates the use of dual problem solutions to directly infer properties of the primal problem's solution.

- It enables the derivation of optimality conditions, such as the Karush-Kuhn-Tucker (KKT) conditions, which provide necessary and sufficient conditions for optimality in constrained optimization problems.

- It facilitates computational strategies that solve the dual problem as a means to solving the primal problem, often leading to more efficient algorithms.

4.4   Karush-Kuhn-Tucker (KKT) Conditions

The Karush-Kuhn-Tucker (KKT) conditions are necessary conditions for a solution to be optimal in a constrained optimization problem. For certain problems, including convex optimization problems that satisfy Slater's condition, they are also sufficient for optimality. The KKT conditions apply to problems of the form:

Minimize $f(x)$ subject to $f_i(x) \leq 0$ for $i = 1, \ldots, M$, and $h_j(x) = 0$ for $j = 1, \ldots, P$, where $f(x)$ is the objective function, $f_i(x)$ are inequality constraints, and $h_j(x)$ are equality constraints.

The KKT conditions consist of the following:

1. **Stationarity:**

$$\nabla f(x) + \sum_{i=1}^{M} \lambda_i \nabla f_i(x) + \sum_{j=1}^{P} \nu_j \nabla h_j(x) = 0,$$

   where λ_i and ν_j are the Lagrange multipliers for the inequality and equality constraints, respectively.

2. **Primal Feasibility:**

$$f_i(x) \leq 0, \quad \forall i = 1, \ldots, M, \quad \text{and} \quad h_j(x) = 0, \quad \forall j = 1, \ldots, P.$$

3. **Dual Feasibility:**
$$\lambda_i \geq 0, \quad \forall i = 1, \ldots, M.$$

4. **Complementary Slackness:**

$$\lambda_i f_i(x) = 0, \quad \forall i = 1, \ldots, M.$$

   This condition implies that for each inequality constraint, either the constraint is active (i.e., $f_i(x) = 0$) and the corresponding Lagrange multiplier λ_i may be positive, or the constraint is not active (i.e., $f_i(x) < 0$) and the corresponding Lagrange multiplier must be zero. The complementary slackness is basically what is used in the case of SVMs to obtain the bias term given the dual solution.

The KKT conditions are a powerful tool in optimization for several reasons:

- They provide a way to check whether a given point is a local optimum for problems with inequality constraints.

- In the context of convex optimization problems that satisfy Slater's condition, the KKT conditions are not only necessary but also sufficient for global optimality.

- The KKT conditions form the basis for many optimization algorithms, especially those designed for nonlinear programming.

Understanding and applying the KKT conditions is fundamental in the field of optimization, particularly for problems involving constraints. They offer a comprehensive set of criteria that must be satisfied at an optimal solution, bridging the gap between theory and practical algorithm design.

4.5 Concluding Remarks

The concepts of weak and strong duality are foundational in optimization theory, bridging primal and dual problems and enabling a deeper understanding of problem structures. For practitioners, recognizing when strong duality holds is key to leveraging dual problems for efficient optimization, especially in the context of convex optimization and machine learning applications such as support vector machines.

5 Application to Linear SVMs

The optimization problem for linear SVMs can be formulated as a primal problem, and its dual is derived using Lagrangians. The dual problem of SVMs is particularly interesting because it only involves the Lagrange multipliers, and the solution to the dual problem can be used to compute the primal solution (w, b).

5.1 Deriving the Dual for Linear SVMs

The primal problem for a linear SVM can be formulated as follows:

Minimize with respect to w and b:

$$\frac{1}{2}\|w\|^2$$

Subject to the constraints for all $i = 1, \ldots, m$:

$$y^{(i)}(w^T x^{(i)} + b) \geq 1,$$

where w is the weight vector, b is the bias term, $x^{(i)}$ are the input features, and $y^{(i)}$ are the labels.

5.1.1 Forming the Lagrangian

The Lagrangian for this problem, incorporating Lagrange multipliers α_i for each constraint, is given by:

$$L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{m} \alpha_i \left(y^{(i)}(w^T x^{(i)} + b) - 1 \right),$$

where $\alpha_i \geq 0$ are the Lagrange multipliers.

5.1.2 Deriving the Dual Problem

To derive the dual problem, we first find the conditions for optimality by taking the partial derivatives of L with respect to w and b, and setting them to zero:

$$\nabla_w L = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)} = 0 \implies w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)},$$

$$\frac{\partial L}{\partial b} = -\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

Substituting these conditions back into the Lagrangian gives the dual formulation, which depends only on the Lagrange multipliers α:

$$\max_{\alpha} g(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} x^{(i)T} x^{(j)},$$

subject to:

$$\alpha_i \geq 0, \quad \forall i,$$

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

5.1.3 Implications of the Dual Problem

The dual problem highlights several key properties of SVMs:

- The solution depends only on the dot products of input vectors, which allows for the kernel trick to be applied in non-linear cases.

- Only support vectors, for which $\alpha_i > 0$, contribute to the decision boundary, making SVMs sparse and efficient models.

- The dual formulation facilitates the application of convex optimization techniques, even for non-linearly separable data.

5.2 Sequential Minimal Optimization (SMO) Algorithm

Sequential Minimal Optimization (SMO) is an algorithm for solving the optimization problem of support vector machines (SVMs) efficiently. SMO breaks down the overall problem into 2-dimensional sub-problems that can be solved analytically, eliminating the need for a numerical optimization algorithm. For more details on the algorithm, we encourage the readers to read the Wikipedia article (`https://en.wikipedia.org/wiki/Sequential_minimal_optimization`) and the original SMO paper (`https://www.microsoft.com/en-us/research/uploads/prod/1998/04/sequential-minimal-optimization.pdf`.

5.3 Obtaining the Primal Solution from the Dual

Once the dual problem of an SVM has been solved, we obtain a set of Lagrange multipliers α_i. These multipliers are instrumental in constructing the primal solution, namely the weight vector w and the bias term b, which define the decision boundary of the SVM.

5.3.1 Computing the Weight Vector w

The weight vector w can be computed directly from the Lagrange multipliers and the training data. Recall the relationship between w, α_i, and the training examples from the formulation of the Lagrangian:

$$w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)},$$

where $x^{(i)}$ are the input features, $y^{(i)}$ are the labels, and m is the number of training examples. Only the support vectors, for which $\alpha_i > 0$, contribute to the sum, making w a linear combination of the support vectors weighted by their corresponding α_i values and labels.

5.3.2 Determining the Bias Term b

The bias term b can be determined using the complementary slackness. Essentially, note that for all points such that $\alpha_j > 0$ (which in the case of SVM are basically the support vectors), the bias term b satisfies:

$$y^{(j)} - w^T x^{(j)} = b.$$

To improve numerical stability, it is common to average over all support vectors on the margin:

$$b = \frac{1}{|S|} \sum_{j \in S} \left(y^{(j)} - w^T x^{(j)} \right),$$

where S is the set of indices of support vectors on the margin, and $|S|$ is the number of such support vectors.

5.3.3 Using w and b for Prediction

With w and b computed, the decision function for a new input x is given by:

$$f(x) = \text{sign}(w^T x + b).$$

This function classifies new examples based on their location relative to the decision boundary defined by w and b.

6 Conclusion

Duality and Lagrangians provide powerful tools for understanding and solving optimization problems. In the context of SVMs, they offer an elegant way to derive solutions that maximize the margin between classes, showcasing the deep connection between primal and dual optimization problems.