1 A Brief Introduction to Duality and KKT Conditions

In the lecture, we explored the derivation of kernel SVM. Essentially, we transformed the primal optimization problem, which aims to maximize the margin, into a dual problem that incorporates the Lagrangian and the Lagrange multipliers α_i. Subsequently, we asserted that we can determine the optimal solution to the dual problem by identifying values of α_{ii} that satisfy the Karush-Kuhn-Tucker (KKT) conditions. In this problem, we will delve more deeply into this process, particularly focusing on the conceptual basis of the dual optimization problem and the relationship between the KKT conditions and optimality.

Optimization problem. Consider an optimization problem in its standard form

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & f_0(\mathbf{x}) \\ \text{subject to} \quad & f_i(\mathbf{x}) \leq 0 \quad \text{for all } i \in \{1, \ldots, m\} \\ & g_i(\mathbf{x}) = 0 \quad \text{for all } i \in \{1, \ldots, n\}. \end{aligned}$$

Denote p^* to be the optimal value ($\min f_0(\mathbf{x})$) to this *primal* problem. For any given optimization problem, we can consider the Lagrangian, which you can treat it as just an expression that we pull out from nowhere for now:

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{n} \nu_i g_i(\mathbf{x})$$

where λ_i and ν_i are known as the Lagrange multipliers associated f_i and g_i respectively. Let $\boldsymbol{\lambda} = (\lambda_i)_i$ and $\boldsymbol{\nu} = (\nu_i)_i$. Cruising without motivation once again, consider the Lagrange dual function

$$\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \inf_{\mathbf{x}} \left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i f_i(\mathbf{x}) + \sum_{i=1}^{n} \nu_i g_i(\mathbf{x}) \right).$$

If you are not familiar with the inf operator, you can treat it as min for now.

Problem 1. Write the SVM primal problem in the standard form. That is, identify the objective f_0 and the constraints $\{f_i\}_i$ and $\{g_i\}_i$ (if any) that are associated to the SVM primal problem.

Problem 2. Write the Lagrangian and Lagrange dual function associated to the SVM primal problem.

Lower bound property. The Lagrange dual function has a fascinating property: if $\lambda_i \geq 0$ for all i, then it is always a lower bound for the primal problem. We shall write $\boldsymbol{\lambda} \geq 0$ as a shorthand for $\lambda_i \geq 0$ for all $i \in \{1, \ldots, m\}$. This lower bound property can then be more formally stated as: for any $\boldsymbol{\lambda} \geq 0$ and $\boldsymbol{\nu}$ (no restriction on what $\boldsymbol{\nu}$ can be), we have $\mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\nu}) \leq p^*$. To see this, let $\tilde{\mathbf{x}}$ be feasible (i.e. it satisfy all constraints, but it is not necessary optimal) and $\boldsymbol{\lambda} \geq 0$. Then

$$f_0(\tilde{\mathbf{x}}) \geq L(\tilde{\mathbf{x}}, \boldsymbol{\lambda}, \boldsymbol{\nu}) \geq \inf_{\mathbf{x}} L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\nu}) = \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\nu}).$$

The first inequality holds because $\tilde{\mathbf{x}}$ is feasible; the second inequality holds by definition of infimum (again, you can treat this as minimum for now if you are not familiar with the subtlety involved with the use of infimum). Since this chain of inequalities applies for any choice of $\tilde{\mathbf{x}}$, if we choose $\tilde{\mathbf{x}}$ to be the one that minimizes $f_0(\mathbf{x})$, we have $f_0(\tilde{\mathbf{x}}) = p^*$, which implies $p^* \geq \mathcal{L}(\boldsymbol{\lambda}, \boldsymbol{\nu})$ as long as $\boldsymbol{\lambda} \geq 0$.

Dual problem. Motivated by the lower bound property, let us now consider the following optimization problem that is induced by the original *primal* problem

$$\begin{aligned} \underset{\boldsymbol{\lambda},\boldsymbol{\nu}}{\text{maximize}} \quad & \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\nu}) \\ \text{subject to} \quad & \lambda_i \geq 0 \qquad \text{for all } i \in \{1,\ldots,m\}. \end{aligned}$$

This is known as the *dual* problem. Note that since $\mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\nu})$ is a lower bound for p^* when $\boldsymbol{\lambda} \geq 0$, this means that any feasible $\boldsymbol{\lambda},\boldsymbol{\nu}$ to this dual problem (i.e. $\boldsymbol{\lambda},\boldsymbol{\nu}$ need not be optimal) is a lower bound of p^*. Intuitively, this means that by solving the *dual* problem, we are finding the tightest lower bound for p^*.

Denote d^* to be the optimal value ($\max \mathcal{L}(\boldsymbol{\lambda},\boldsymbol{\nu})$) to this *dual* problem. Due to the lower bound property, we already know that $p^* \geq d^*$. This is known as *weak duality*. However, under some nice condition[1], we could have $p^* = d^*$. This is known as *strong duality*.

Problem 3. Write down the SVM dual problem.

Complementary slackness. Assume that strong duality holds (i.e. $p^* = d^*$), with $\mathbf{x}^*$ be primal optimal and $\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*$ be dual optimal. Observe that

$$\begin{aligned} \overbrace{f_0(\mathbf{x}^*)}^{p^*} = \overbrace{\mathcal{L}(\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)}^{d^*} &= \inf_{\mathbf{x}} \overbrace{\left(f_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}) + \sum_{i=1}^{n} \nu_i^* g_i(\mathbf{x}) \right)}^{L(\mathbf{x},\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)} \\ &\leq f_0(\mathbf{x}^*) + \sum_{i=1}^{m} \lambda_i^* f_i(\mathbf{x}^*) + \sum_{i=1}^{n} \nu_i^* g_i(\mathbf{x}^*) \\ &\leq f_0(\mathbf{x}^*). \end{aligned}$$

Since the chain of inequalities are sandwiched by $f_0(\mathbf{x}^*)$, we know that the two inequalities holds with equality. This first inequality (turned equality) implies that $\mathbf{x}^*$ minimizes $L(\mathbf{x},\boldsymbol{\lambda}^*,\boldsymbol{\nu}^*)$. More interestingly, the second inequality (turned equality) implies that $\lambda_i^* f_i(\mathbf{x}^*) = 0$ for all $i \in \{1,\ldots,m\}$. This is also known as *complementary slackness*, which is aptly named after the fact that if λ_i^* is slacked (i.e. $\lambda_i^* > 0$), then $f_i(\mathbf{x}^*)$ is tight (i.e. $f_i(\mathbf{x}^*) = 0$), and vice versa.

Problem 4. What is the interpretation of complementary slackness in SVM?

Karush-Kuhn-Tucker (KKT) conditions. The KKT conditions is a set of conditions that a tuple of primal and dual variable $(\tilde{\mathbf{x}},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{\nu}})$ might (or might not) hold. In particular, the KKT conditions are

1. Primal feasibility: $f_i(\tilde{\mathbf{x}}) \leq 0$ and $g_i(\tilde{\mathbf{x}}) = 0$.
2. Dual feasibility: $\tilde{\boldsymbol{\lambda}} \geq 0$.
3. Complementary slackness: $\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0$.
4. Vanishing gradient: $\nabla_{\mathbf{x}} L(\mathbf{x},\tilde{\boldsymbol{\lambda}},\tilde{\boldsymbol{\nu}}) = 0$ at $\mathbf{x} = \tilde{\mathbf{x}}$.

[1] An example of such nice condition is to have convex f_0 and f_i, have affine g_i, and satisfy a mild condition known as Slater's condition. It is not necessary to understand this deeply for now, but for the curious, refer to its Wikipedia page to see the complete statement of Slater's condition.

A common misconception revolving the use of KKT conditions is: if we found some $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ that satisfy these conditions, then they are optimal. This is not always true, and here is a counterexample: Consider the optimization problem minimize $-x^2$ with no constraints. Then condition (1)-(3) hold vacuously for $\tilde{x} = 0$, and we can easily verify that condition (4) holds for $\tilde{x}$ too. However, $\tilde{x} = 0$ is not optimal even though it satisfied all of KKT conditions.

KKT and optimality. The natural question is: what is the relationship between optimality and the KKT conditions? Here are two facts:

1. If strong duality holds and $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is optimal, then the KKT conditions must hold for $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$.
2. Let f_i be convex for all $i \in \{0, \dots, m\}$ and g_i be affine for all $i \in \{1, \dots, n\}$. If KKT conditions hold for $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$, then $(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ is optimal with strong duality.

The first statement is straightforward, and you should try to prove it yourself to test your understanding so far. To prove the second statement, observe that

$$\begin{aligned} L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) &= f_0(\tilde{\mathbf{x}}) + \sum_{i=1}^{m} \tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) + \sum_{i=1}^{n} \tilde{\nu}_i g_i(\tilde{\mathbf{x}}) \\ &= f_0(\tilde{\mathbf{x}}) \end{aligned}$$

because $\tilde{\lambda}_i f_i(\tilde{\mathbf{x}}) = 0$ from complementary slackness and $g_i(\tilde{\mathbf{x}}) = 0$ from primal feasibility. Furthermore, $\mathcal{L}(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) = L(\tilde{\mathbf{x}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ because the $\nabla_{\mathbf{x}} L$ at $\tilde{\mathbf{x}}$ is zero by vanishing gradient (which implies that $\tilde{\mathbf{x}}$ is a local minimum for L) and L is convex (which implies the local minimum is also a global minimum). Since both $f_0(\tilde{\mathbf{x}})$ and $\mathcal{L}(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}})$ equals to the same quantity, we have $p^* \leq f_0(\tilde{\mathbf{x}}) = \mathcal{L}(\tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{\nu}}) \leq d^* \leq p^*$, which implies optimality with strong duality because the inequalities are sandwiched by p^*.

Bringing it back to SVM. In the primal formulation of SVM, we see that the objective $\|\mathbf{w}\|^2/2$ is convex in $\mathbf{w}$ and the inequality constraint $1 - t_i(\mathbf{w}^\top \mathbf{x}_i + b)$ is affine in $\mathbf{w}$ (which implies convexity). As such, by the second statement, if the KKT conditions hold for variables $(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\alpha}})$, then $(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\alpha}})$ is optimal. In other words, to find the optimal weights $\mathbf{w}^*$, all we have to do is to find $(\tilde{\mathbf{w}}, \tilde{\boldsymbol{\alpha}})$ that satisfy the KKT conditions and let $\mathbf{w}^* = \tilde{\mathbf{w}}$.

2 SVM for Non-Separable Data

The SVM that we formulated in lecture assumes that the data are separable. This assumption is required to ensure that the constraints $t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1$ are satisfied for all $n \in \{1, \ldots, N\}$. In this question, we shall extend SVM to allow for non-separable data.

The idea is simple: we modify our constraints to accommodate for the non-separability. Instead of insisting that all data points remain on one side of their respective hyperplane, we relax this constraint, permitting certain points to cross over their associated hyperplane. More formally, we "slacken" the constraint $t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1$ to $t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 - \xi_n$. Here, $\xi_n \geq 0$ is known as the slack variable for data point $\mathbf{x}_n$. The larger ξ_n is, the further the $\mathbf{x}_n$ crosses away from its hyperplane. As such, we want to penalize large ξ_n. This lead us to the following optimization problem

$$\begin{aligned} \underset{\mathbf{w}, b, \{\xi_n\}_n}{\text{minimize}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=1}^{N} \xi_n \\ \text{subject to} \quad & t_n(\mathbf{w}^\top \mathbf{x}_n + b) \geq 1 - \xi_n \quad \text{for all } n \in \{1, \ldots, N\} \\ & \xi_n \geq 0 \quad \text{for all } n \in \{1, \ldots, N\}. \end{aligned}$$

Problem 1. Let $\boldsymbol{\alpha} = \{\alpha_n\}_n$ and $\boldsymbol{\lambda} = \{\lambda_n\}_n$ be the Lagrange multiplier associated to the first and second set of constraints. Write the Lagrangian $L(\mathbf{w}, b, \{\xi_n\}_n, \boldsymbol{\alpha}, \boldsymbol{\lambda})$ for the optimization problem.

Problem 2. What constraints must be imposed on $\boldsymbol{\alpha}$ and $\boldsymbol{\lambda}$?

Finding the Lagrange dual function. Recall that the Lagrange dual function

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \min_{\mathbf{w}, b, \{\xi_n\}_n} L(\mathbf{w}, \{\alpha_n\}_n, \{\lambda_n\}_n).$$

In other words, to find $\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda})$, we need to minimize the Lagrangian L with respect to $\mathbf{w}$, b and $\{\xi_n\}_n$.

Problem 3. Find $\nabla_{\mathbf{w}} L$, $\partial L/\partial b$, and $\partial L/\partial \xi_n$.

Problem 4. By setting the derivatives found in Problem 3 to zero, what constraints are revealed?

Problem 5. By substituting the constraint associated to $\mathbf{w}$ from Problem 4, verify that the Lagrangian $L(\mathbf{w}, b, \{\xi_n\}_n, \boldsymbol{\alpha}, \boldsymbol{\lambda})$ can be expressed as

$$\sum_{n=1}^{N}(C - \alpha_n - \lambda_n)\xi_n + \frac{1}{2}\left\|\sum_{n=1}^{N}\alpha_n t_n \mathbf{x}_n\right\|^2 + \sum_{n=1}^{N}\alpha_n - \left(\sum_{n=1}^{N}\alpha_n t_n\right) b - \sum_{n=1}^{N}\alpha_n t_n \left(\sum_{m=1}^{N} t_m \alpha_m \mathbf{x}_m\right)^\top \mathbf{x}_n.$$

Problem 6. Using the remaining constraints found in Problem 4, verify that the Lagrange dual function

$$\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\lambda}) = \sum_{n=1}^{N}\alpha_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N}\alpha_n \alpha_m t_n t_m \mathbf{x}_n^\top \mathbf{x}_m.$$

Problem 7. Observe that $\boldsymbol{\lambda}$ does not appear in $\mathcal{L}$. As such, we can try to combine some of the constraints to simplify the dual problem. After performing this simplification, verify that the only modification required for allowing non-separable data is the introduction of the constraint $\alpha_n \leq C$.