Support Vector Machines Lecture Supplement

The Lagrangian

Given the minimization optimization problem

$$\min_{\overline{x}} \phi(\overline{x}),$$

subject to

$$g_i(\overline{x}) \geq 0,$$

forall $i = 1, \ldots, l$, the **Lagrangian** of this optimization problem is defined as

$$L(\overline{\alpha}, \overline{x}) = \max_{\overline{\alpha}} \min_{\overline{x}} \phi(\overline{x}) - \sum_{i=1}^{l} \alpha_i g_i(\overline{x}),$$

where $\alpha_i \geq 0$, $i = 1, \ldots, l$.

Notice how the Lagrangian represents nested optimization problems, where the outer problem is a maximization problem over $\overline{\alpha}$, and the inner problem is a minimization problem over $\overline{x}$, but that also depends on the current value of $\overline{\alpha}$, which is treated as a constant within the inner optimization problem. The optimal solution to the Lagrangian is the $(\overline{\alpha}^*, \overline{x}^*)$ for which $L(\overline{\alpha}^*, \overline{x}^*)$ is a maximum over all such tuples.

The following algorithm demonstrates the logic of how $(\overline{\alpha}^*, \overline{x}^*)$ is determined.

The Lagrangian Algorithm

Input objective function ϕ and constraints $g_1, \ldots, g_l$.

Initialize max: max $\leftarrow -\infty$.

forall $\overline{\alpha} = (\alpha_1, \ldots, \alpha_l) \geq \overline{0}$

Initialize min: min $\leftarrow \infty$

forall $\overline{x}$

$m \leftarrow \phi(\overline{x}) - \sum_{i=1}^{l} \alpha_i g_i(\overline{x})$.

If $m <$ min

min $\leftarrow m$.

$\overline{x}_{\min} \leftarrow \overline{x}$.

If min $>$ max

max $\leftarrow$ min.

$\overline{\alpha}^* \leftarrow \overline{\alpha}$.

$\overline{x}^* \leftarrow \overline{x}_{\min}$.

Return $L(\overline{\alpha}^*, \overline{x}^*)$.

Thus, the algorithm searches for an $\overline{\alpha}$ that makes the inner optimization problem result in a value that is as large as possible.

Another way of thinking of the Lagrangian is that of a battle between adversaries `Max` and `Min`. `Max` is searching for an $\overline{\alpha}$ that makes the inner optimization problem result in a value that is as *large* as possible. But for each $\overline{\alpha}$ that she passes to `Min`, `Min` attempts to produce a value that is as *small* as possible.

For example, one strategy that `Max` might try is to assign $\overline{\alpha} = \overline{0}$. This has the positive effect of preventing `Min` from attaining a small value by inducing some $g_i(\overline{x})$ to become a large positive value. On the other hand, setting $\overline{\alpha} = \overline{0}$ eliminates the g constraints from the inner objective function, and thus allows `Min` to make ϕ as small as possible without incurring any constraint penalties, which could be detrimental to `Max`.

As another example, suppose `Max` assigns some α_i to a large value. Then `Min` will make sure to satisfy constraint $g_i(\overline{x}) \geq 0$ so that $-\alpha_i g_i(\overline{x})$ will make for a large negative value, again hurting `Max`. Thus, `Max` is searching for an $\overline{\alpha}$ that minimizes the damage inflicted by `Min`; i.e. forces `Min` to return the largest minimum value as possible.

Karush Kuhn Tucker (KKT) Conditions

It should be emphasized that the purpose of the Lagrangian dual formulation is to solve the original primal optimization problem. Moreover, the Karush Kuhn Tucker (KKT) conditions provide necessary and sufficient conditions for when, in the optimal Lagrangian solution $(\overline{\alpha}^*, \overline{x}^*)$, $\overline{x}^*$ is the optimal solution to the primal problem.

Theorem 1 (Karush Kuhn Tucker). Suppose the Lagrangian $L(\overline{\alpha}, \overline{x})$ is a differentiable function with optimal solution $(\overline{\alpha}^*, \overline{x}^*)$. Then $\overline{x}^*$ is the optimal solution to the primal problem iff each of the following conditions hold.

1. $\frac{\partial L(\overline{\alpha}^*, \overline{x}^*)}{\partial \overline{x}} = 0$
2. $g_i(\overline{x}^*) \geq 0$, for all $i = 1, \ldots, l$
3. $\alpha_i^* g_i(\overline{x}^*) = 0$, for all $i = 1, \ldots, l$
4. $\overline{\alpha}^* \geq \overline{0}$

Prove of Theorem 1. First notice that the Condition 4 always holds by definition of the Largrangian. Also, since $(\overline{\alpha}^*, \overline{x}^*)$ is an optimal solution to the Lagrangian and L is differentiable, the fact that

$$\frac{\partial L(\overline{\alpha}^*, \overline{x}^*)}{\partial \overline{x}} = 0$$

is a result from the calculus of functions. Hence Condition 1 holds for any optimal solution $(\overline{\alpha}^*, \overline{x}^*)$.

Next suppose that, in the optimal Lagrangian solution $(\overline{\alpha}^*, \overline{x}^*)$, $\overline{x}^*$ is the optimal solution to the primal problem. Then, since $\overline{x}^*$ is the optimal primal solution, it must satisfy the constraints $g_i(\overline{x}^*) \geq 0$, for all $i = 1, \ldots, l$. See Exercise 1 for the establishment of Condition 3.

Conversely, now assume Conditions 2 and 3 hold with respect to optimal Lagrangian solution $(\overline{\alpha}^*, \overline{x}^*)$. Then from Condition 2, $\overline{x}^*$ is a feasible solution to the primal optimization problem. Moreover, by Condition 3, $L(\overline{\alpha}^*, \overline{x}^*) = \phi(\overline{x}^*)$. Now suppose by way of contradiction that $\overline{x}^*$ is not the optimal primal solution, and let $\overline{y}^*$ be a primal solution for which $\phi(\overline{y}^*) < \phi(\overline{x}^*)$. Then $g_i(\overline{y}^*) \geq 0$, for all $i = 1, \ldots, l$, and we may define $\overline{\gamma} \geq \overline{0}$ so that $\gamma_i g_i(\overline{y}^*) = 0$, for all $i = 1, \ldots, l$. But then,

$$L(\overline{\gamma}, \overline{y}^*) = \phi(\overline{y}^*) < L(\overline{\alpha}^*, \overline{x}^*) = \phi(\overline{x}^*),$$

which contradicts the optimality of $(\overline{\alpha}^*, \overline{x}^*)$. Therefore, $\overline{x}^*$ is the optimal solution to the primal problem.

Note that it can be shown that, in the special case when ϕ is i) differentiable and convex, and ii) the g_i constraints are linear, then the four KKT conditions will hold for an optimal solution to the Lagrangian. This is good news for the theory of support vector machines, since the the primal problem for finding a maximum-margin linear classifier satisfies properties i) and ii).

Dual Lagrangian Examples

Example 1. Solve the optimization problem of minimizing $\phi(x) = 6 - \sqrt{x}$ subject to $x \leq 4$.

Example 1 Solution.

Example 2. Compute the Lagrange dual for the optimization problem of minimizing $\phi(x, y) = x^2+y^2$ subject to $x \geq 0$, $y \geq 0$, and $x + y \geq 1$.

Example 2 Solution.

The Lagrangian Dual for the Maximum-Margin Optimization

Let $f(\overline{w})$ be a real-value function that takes as input a real-valued vector $\overline{w} = w_1, \ldots, w_n$. Then

$$\frac{\partial f(\overline{w})}{\partial \overline{w}} = \left(\frac{\partial f(\overline{w})}{\partial w_1}, \ldots, \frac{\partial f(\overline{w})}{\partial w_n} \right)$$

is defined as an n-dimensional vector whose i th component is the partial derivative of f with respect to variable w_i, $i = 1, \ldots, n$.

To diifferentiate the Lagrangian of the maximum-margin optimization problem, we need the following result.

Theorem 2. The following derivative formulas hold.

a. If $\overline{c}$ is a constant vector, then
$$\frac{\partial(\overline{c}\cdot\overline{w})}{\partial\overline{w}} = \overline{c}.$$

b.
$$\frac{\partial(\overline{w}\cdot\overline{w})}{\partial\overline{w}} = 2\overline{w}.$$

We prove 1a. and leave 2a. as an exercise. By definition,

$$\overline{c}\cdot\overline{w} = c_1w_1 + \cdots + c_nw_n.$$

Then, for all $i = 1, \ldots, n$,
$$\frac{\partial(\overline{c}\cdot\overline{w})}{\partial w_i} = c_i.$$

Therefore,
$$\frac{\partial(\overline{c}\cdot\overline{w})}{\partial\overline{w}} = (c_1, \ldots, c_n) = \overline{c}.$$

Corollary 1. Given the Lagrangian

$$L(\overline{\alpha}, \overline{w}, b) = \frac{1}{2}\overline{w}\cdot\overline{w} - \sum_{i=1}^{l}(\alpha_i y_i(\overline{w}\cdot\overline{x}_i - b) - \alpha_i),$$

KKT Condition 1 implies the following.

$$\text{a. } \overline{w}^* = \sum_{i=1}^{l}\alpha_i^* y_i\overline{x}_i.$$

$$\text{b. } \sum_{i=1}^{l}\alpha_i^* y_i = 0.$$

Proof of Corollary 1. KKT Condition 1 implies $\frac{\partial L(\overline{\alpha}^*, \overline{w}, b)}{\partial\overline{w}} = 0$. Then by Theorem 2,

$$\frac{\partial L(\overline{\alpha}^*, \overline{w}, b)}{\partial\overline{w}} = \overline{w} - \sum_{i=1}^{l}\alpha_i y_i x_i = 0,$$

when $\overline{w} = \overline{w}^*$. Hence,

$$\overline{w}^* = \sum_{i=1}^{l}\alpha_i^* y_i\overline{x}_i.$$

Also, KKT Condition 1 implies $\frac{\partial L(\overline{\alpha}^*, \overline{w}, b)}{\partial b} = 0$, which implies

$$\sum_{i=1}^{l} \alpha_i^* y_i = 0,$$

since the b variable is mulitplied by $\alpha_1^* y_1 + \cdots + \alpha_l^* y_l$.

Theorem 3. The Lagrange dual for the MMC optimization problem is

$$\phi'(\overline{\alpha}) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\overline{x}_i \cdot \overline{x}_j).$$

Proof of Theorem 3. Substituting for $\overline{w}$ using the KKT-Condition-1 constraint

$$\overline{w} = \sum_{i=1}^{l} \alpha_i y_i \overline{x}_i,$$

and using the KKT-Condition-1 constraint

$$\sum_{i=1}^{l} \alpha_i y_i = 0,$$

we have

$$\phi'(\overline{\alpha}) = \frac{1}{2} \overline{w} \cdot \overline{w} - \sum_{i=1}^{l} (\alpha_i y_i (\overline{w} \cdot \overline{x}_i - b) - \alpha_i) =$$

$$\frac{1}{2} (\sum_{i=1}^{l} \alpha_i y_i \overline{x}_i) \cdot (\sum_{j=1}^{l} \alpha_j y_j \overline{x}_j) - \sum_{i=1}^{l} \alpha_i y_i (\sum_{j=1}^{l} \alpha_j y_j \overline{x}_j) \cdot \overline{x}_i + b \sum_{i=1}^{l} \alpha_i y_i + \sum_{i=1}^{l} \alpha_i =$$

$$\frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\overline{x}_i \cdot \overline{x}_j) - \sum_{i=1}^{l} \alpha_i y_i \overline{x}_i \cdot (\sum_{j=1}^{l} \alpha_j y_j \overline{x}_j) + 0 + \sum_{i=1}^{l} \alpha_i =$$

$$\sum_{i=1}^{l} \alpha_i + \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\overline{x}_i \cdot \overline{x}_j) - \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\overline{x}_i \cdot \overline{x}_j) =$$

$$\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j (\overline{x}_i \cdot \overline{x}_j).$$

Canonical Dot Product Spaces

Let $k : \mathcal{R}^n \times \mathcal{R}^n \to \mathcal{R}$ be a kernel. Then k satisfies the following properties.

Symmetry $k(\overline{x}, \overline{y}) = k(\overline{y}, \overline{x})$.

Positive Definite For any m scalars $\theta_1, \ldots, \theta_m$ and m vectors $\overline{x}_1, \ldots, \overline{x}_m$,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \theta_i \theta_j k(\overline{x}_i, \overline{x}_j) \geq 0.$$

We now define a function Φ that maps each vector $\overline{x} \in \mathcal{R}^n$ to a vector $\Phi(\overline{x})$ in a higher-dimensional dot product space, in such a way that

$$k(\overline{x}, \overline{y}) = \Phi(\overline{x}) \cdot \Phi(\overline{y}).$$

The mapping Φ is defined as $\Phi(\overline{x}) = k(, \overline{x})$, which is a function f from $\mathcal{R}^n$ to $\mathcal{R}$, with the property that $f(\overline{y}) = k(\overline{y}, \overline{x})$. Now since two functions can be added to make another function, and a scalar times a function results in another function, we may think of the set of functions that are linear combinations of functions of the form $k(, \overline{x})$ as a vector space (it is an exercise to verify that all the vector-space axioms are satisfied). Thus, an element of this space has the form

$$\theta_1 k(, \overline{x}_1) + \cdots + \theta_m k(, \overline{x}_m),$$

where $\theta_1, \ldots, \theta_m$ are scalars, and $\overline{x}_1, \ldots, \overline{x}_m \in \mathcal{R}^n$. Now given two vectors $f = \theta_1 k(, \overline{x}_1) + \cdots + \theta_m k(, \overline{x}_m)$ and $g = \gamma_1 k(, \overline{y}_1) + \cdots + \gamma_p k(, \overline{y}_p)$, we define the dot product between f and g as

$$f \cdot g = \sum_{i=1}^{m} \sum_{j=1}^{p} \theta_i \gamma_j k(\overline{x}_i, \overline{y}_j).$$

To verify that this is a dot product, we must establish the following properties.

Symmetry $f \cdot g = g \cdot f$

Additivity $f \cdot (g + h) = f \cdot g + f \cdot h$

Scalar Associativity $(\alpha f) \cdot g = \alpha(f \cdot g)$, for all real α

Positivity $f \cdot f \geq 0$ and $f \cdot f = 0$ iff $f = \overline{0}$

Symmetry follows from the symmetry of k. To show scalar associativity,

$$(\alpha f) \cdot g = \sum_{i=1}^{m} \sum_{j=1}^{p} (\alpha \theta_i) \gamma_j k(\overline{x}_i, \overline{y}_j) =$$

$$\sum_{i=1}^{m}\sum_{j=1}^{p}\alpha(\theta_i\gamma_j)k(\overline{x}_i,\overline{y}_j)=\alpha\sum_{i=1}^{m}\sum_{j=1}^{p}\theta_i\gamma_j k(\overline{x}_i,\overline{y}_j)=\alpha(f\cdot g).$$

To show additivity, assume $g=\gamma_1 k(,\overline{y}_1)+\cdots+\gamma_l k(,\overline{y}_l)$ and $h=\gamma_{l+1}k(,\overline{y}_{l+1})+\cdots+\gamma_p k(,\overline{y}_p)$. Then

$$f\cdot(g+h)=\sum_{i=1}^{m}\sum_{j=1}^{p}\theta_i\gamma_j k(\overline{x}_i,\overline{y}_j)=$$

$$\sum_{i=1}^{m}\sum_{j=1}^{l}\theta_i\gamma_j k(\overline{x}_i,\overline{y}_j)+\sum_{i=1}^{m}\sum_{j=l+1}^{p}\theta_i\gamma_j k(\overline{x}_i,\overline{y}_j)=f\cdot g+f\cdot h.$$

Finally, to show positivity, if $f=\theta_1 k(,\overline{x}_1)+\cdots+\theta_m k(,\overline{x}_m)$, then

$$f\cdot f=\sum_{i=1}^{m}\sum_{j=1}^{m}\theta_i\theta_j k(\overline{x}_i,\overline{x}_j)\geq 0,$$

by the positive-definiteness property of k.

Now assume $f\cdot f=0$. We must prove that $f=0$, meaning that $f(\overline{x})=0$ for all inputs $\overline{x}\in\mathcal{R}^n$. This is equivalent to proving that $|f(\overline{x})|^2=0$ for all $\overline{x}\in\mathcal{R}^n$. To prove this we need the Cauchy-Schwarz inequality. We now give a purely algebraic proof of this inequality that only relies on the above properties of the dot product (except for the second part of the positivity property which we still have yet to establish for our vector space of functions).

Theorem 3 (Cauchy-Schwarz-Bunyakovsky Inequality). If $\overline{u}$ and $\overline{v}$ are vectors in some dot product space, then

$$(\overline{u}\cdot\overline{v})^2\leq(\overline{u}\cdot\overline{u})(\overline{v}\cdot\overline{v}).$$

Proof of Theorem 3. For any scalar t, and by several applications of the addivity, symmetry, and scalar associativity properties , we get

$$0\leq(t\overline{u}+\overline{v})\cdot(t\overline{u}+\overline{v})=t^2(\overline{u}\cdot\overline{u})+2t(\overline{u}\cdot\overline{v})+(\overline{v}\cdot\overline{v}),$$

where the inequality is due to the first part of the positivity property (which has already been shown to hold in our function vector space). Furthermore, this inequality may be written as $at^2+bt+c\geq 0$, where $a=\overline{u}\cdot\overline{u}$, $b=2(\overline{u}\cdot\overline{v})$, and $c=\overline{v}\cdot\overline{v}$. But $at^2+bt+c\geq 0$ implies that the equation $at^2+bt+c=0$ either has no roots, or exactly one root. In other words, we must have

$$b^2-4ac\leq 0,$$

which implies

$$4(\overline{u}\cdot\overline{v})^2\leq 4(\overline{u}\cdot\overline{u})(\overline{v}\cdot\overline{v}),$$

or

$$(\overline{u}\cdot\overline{v})^2\leq(\overline{u}\cdot\overline{u})(\overline{v}\cdot\overline{v}),$$

which proves the theorem.

Returning to establishing the second part of positivity, assume that $f = \theta_1 k(,\overline{x}_1) + \cdots + \theta_m k(,\overline{x}_m)$ and that $f \cdot f = 0$. Let $\overline{x} \in \mathcal{R}^n$ be arbitrary. Then

$$k(,\overline{x}) \cdot f = \sum_{i=1}^{m} \theta_i k(\overline{x}_i, \overline{x}) = \sum_{i=1}^{m} \theta_i k(\overline{x}, \overline{x}_i) = f(\overline{x}).$$

Hence, by the Cauchy-Schwarz inequality,

$$|f(\overline{x})|^2 = (k(,\overline{x}) \cdot f)^2 \leq (k(,\overline{x}) \cdot k(,\overline{x}))(f \cdot f) = (k(,\overline{x}) \cdot k(,\overline{x})) \cdot 0 = 0.$$

Therefore, $f(\overline{x}) = 0$ for all $\overline{x} \in \mathcal{R}^n$.

Exercises

1. Suppose that optimal Lagrangian solution $(\overline{\alpha}^*, \overline{x}^*)$ is such that $\overline{x}^*$ is the optimal solution to the primal problem of minimizing $\phi(\overline{x})$ subject to $g_i(\overline{x}) \geq 0$, for all $i = 1, \ldots, l$. For each KKT condition, justify why it is satified by $(\overline{\alpha}^*, \overline{x}^*)$.

   a. $\frac{\partial L(\overline{\alpha}^*, \overline{x}^*)}{\partial \overline{x}} = 0$

   b. $g_i(\overline{x}^*) \geq 0$, for all $i = 1, \ldots, l$

   c. $\alpha_i^* g_i(\overline{x}^*) = 0$, for all $i = 1, \ldots, l$

   d. $\overline{\alpha}^* \geq \overline{0}$

2. Prove that if the KKT conditions are satisfied, then $L(\overline{\alpha}^*, \overline{x}^*) = \phi(\overline{x}^*)$, and $\overline{x}^*$ is the solution to the primal problem.

3. Let $\phi(x) = -\ln x$. Suppose ϕ is to be minimized subject to $x \leq e$. Compute the Lagrangian $L(\alpha, x)$ and Lagrange dual $\phi'(\alpha)$ of this optimization problem, and solve the optimization problem by maximizing ϕ'.

4. Let $\phi(x) = x^2 - x$. Suppose ϕ is to be minimized subject to $x \geq 3$. Compute the Lagrangian $L(\alpha, x)$ and Lagrange dual $\phi'(\alpha)$ of this optimization problem, and solve the optimization problem by maximizing ϕ'.

5. Let $\phi(x, y) = 2x^2 + y^2$. Suppose ϕ is to be minimized subject to $x \geq 0$, $y \geq 0$, and $x + 2y \geq 5$. Compute the Lagrangian $L(\alpha, x)$ and Lagrange dual $\phi'(\alpha)$ of this optimization problem.

6. Prove that
$$\frac{\partial(\overline{w} \cdot \overline{w})}{\partial \overline{w}} = 2\overline{w}.$$

7. Recall that to represent an optimization problem as a quadratic program, one requires matrices Q, $\overline{q}$, X, and $\overline{c}$. Moreover these matrices can be used to construct the quadratic program
$$\overline{\alpha}^* = \underset{\overline{\alpha}}{\operatorname{argmax}} \left(\overline{q} \cdot \overline{\alpha} - \frac{1}{2} \overline{\alpha}^t Q \overline{\alpha} \right),$$
   subject to
$$X^t \overline{\alpha} \geq \overline{c}.$$
   Formally define the matrices Q, $\overline{q}$, X, and $\overline{c}$ for solving the Lagrange dual of the MMC optimization problem. Hint: in addition to the constraint $\overline{\alpha} \geq \overline{0}$, don't forget the constraint
$$\sum_{i=1}^{l} \alpha_i y_i = 0.$$

8. Provide the matrices Q, $\overline{q}$, X, and $\overline{c}$ for the Lagrange dual of the MMC optimization problem, for which the positively classified training vectors are
$$(-1, 1), (2, 2), (3, 2),$$
   while the negatively classified training vectors are
$$(-2, -4), (1, -1), \text{ and } (4, -2).$$

9. If positive vectors (x_1, x_2) satisfy $x_1^2 + x_2^2 > 1$, while negative vectors (x_1, x_2) satisfy $x_1^2 + x_2^2 < 1$, then prove that the plane having equation $\overline{w} \cdot \overline{x} = b$, where $\overline{w} = (1, 1, 0)$ and $b = 1$, linearly separates the positive and negative vectors after each is transformed by $\Phi(\overline{x}) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$. Prove that points satisfying $x_1^2 + x_2^2 = 1$ are transformed to points on the plane.

10. Prove that the Cauchy-Schwarz inequality

$$(\overline{u} \cdot \overline{v})^2 \leq (\overline{u} \cdot \overline{u})(\overline{v} \cdot \overline{v})$$

becomes an equality iff $\overline{v}$ is a scalar multiple of $\overline{u}$. Hint: the proof should only use the four axioms of a dot product, and

$$\overline{u} \cdot \overline{v} = u_1v_1 + \cdots + u_nv_n$$

is *not* one of the axioms.

11. Prove that if k is a kernel, then $k(\overline{x}, \overline{x}) \geq 0$, for all $\overline{x} \in \mathcal{R}^n$.

12. Prove that the dot product for $\mathcal{R}^n$ is a kernel.

13. Let $k(x, y) = (\overline{x} \cdot \overline{y})^2$, for all $\overline{x}, \overline{y} \in \mathcal{R}^2$. Consider the canonical mapping Φ from $\mathcal{R}^2$ to the canonical dot product function vector space $\mathcal{F}$ defined by $\Phi(\overline{x}) = k(, \overline{x})$.

    a. Provide a formula for $\Phi((1, 3))$ as a function f of $\overline{x} = (x_1, x_2)$. In other words, give a formula for $f(x_1, x_2)$.

    b. Provide a formula $f(x_1, x_2)$ for the function vector

$$-2\Phi((2, 1)) + 3\Phi((3, 2)) - \Phi((0, 1)).$$

    c. Compute $\Phi((1, 3)) \cdot \Phi((-2, 1))$.

    d. Verify that, for the f from Part a), $f(\overline{x}) = f \cdot k(, \overline{x})$.

14. Prove that the square of the dot product for $\mathcal{R}^n$ is a kernel. Hint: use the previous exercise.

15. Show that the dual Lagrangian for the soft-margin MMC optimization problem is identical to that of the original (hard-margin) version, with the only exception that $\alpha_i \leq C$ for each Lagrange multiplier α_i, $i = 1, \ldots, l$.

Exercise Solutions

1. It helps to think in terms of adversaries Max and Min.

   a. Assuming smoothness of L, for fixed $\overline{\alpha}^*$, $L(\overline{\alpha}^*, \overline{x})$ is a smooth function of $\overline{x}$. Moreover, Min's goal is to minimize this function whenever it receives an $\overline{\alpha}$ value from Max. In case of $\overline{\alpha}^*$, Min accomplished this via $\overline{x} = \overline{x}^*$. Hence, L is a local minimum at point $(\overline{\alpha}^*, \overline{x}^*)$ in the direction of $\overline{x}$. Therefore,
   $$\frac{\partial L(\overline{\alpha}^*, \overline{x}^*)}{\partial \overline{x}} = 0.$$

   b. True since $\overline{x}^*$ is the optimal solution to the primal problem, and it must satisfy all the constraints.

   c. Suppose by way of contradiction that, e.g., $\alpha_1 g_1(\overline{x}^*) \neq 0$. Then from conditions b) and d), we must have $\alpha_1 g_1(\overline{x}^*) > 0$, which means that this term contributes $-\alpha_1 g_1(\overline{x}^*)$ to L. However, since L is smooth and, for fixed $\overline{x}^*$, L has a local maximum at $\overline{\alpha}^*$, it follows that, if we move in the direction of α_1 and slightly reduce its value, then the value of L will actually *increase*, which contradicts that L has a local maximum at $(\overline{\alpha}^*, \overline{x}^*)$ in the direction of $\overline{\alpha}$.

   d. True by the definition of the Lagrangian.

2. By KKT Condition 3, $\alpha_i^* g_i(\overline{x}^*) = 0$, for all $i = 1, \ldots, l$. Hence,
   $$L(\overline{\alpha}^*, \overline{x}^*) = \phi(\overline{x}^*) - 0 = \phi(\overline{x}^*).$$
   Moreover, by KKT Condition 2, all of the constraints are satisfied, and we claim that $\overline{x}^*$ must minimize ϕ subject to these constraints being satisfied. Otherwise, if say $\phi(\overline{y}^*) < \phi(\overline{x}^*)$, then, after Max had played $\overline{\alpha}^*$, Min would have played $\overline{y}^*$ to produce a smaller min value, namely $\phi(\overline{y}^*)$.

3. The constraint $x \leq e$ can be written as $g_1(x) = e - x$. Then $L(\alpha, x) = -\ln x - \alpha(e - x)$. By KKT-Condition 1,
   $$\frac{\partial L(\alpha, x)}{\partial x} = \frac{-1}{x} + \alpha = 0 \Rightarrow x = 1/\alpha.$$
   Then substituting x for $1/\alpha$ gives
   $$\phi'(\alpha) = \ln \alpha - \alpha(e - 1/\alpha).$$
   Notice that $\phi'(\alpha)$ approaches $-\infty$ at the extremes: $x \to 0$ and $x \to \infty$. Thus, its maximum will occur at its unique critical point that satisfies
   $$\frac{d\phi'}{d\alpha} = 1/\alpha - e = 0 \Rightarrow \alpha = 1/e.$$
   Verify that $1/e$ is in fact a local maximum by showing
   $$\frac{d^2\phi'(1/e)}{d\alpha^2} < 0.$$
   Therefore, $x^* = 1/\alpha^* = 1/(1/e) = e$ is the primal solution, which is what we would expect when examining the graph of $-\ln x$.

4. The constraint $x \geq 3$ can be written as $g_1(x) = x - 3$. Then $L(\alpha, x) = x^2 - x - \alpha(x-3)$. By KKT-Condition 1,
$$\frac{\partial L(\alpha, x)}{\partial x} = 2x - 1 - \alpha = 0 \Rightarrow x = (\alpha + 1)/2.$$
Then substituting x for $(\alpha + 1)/2$ gives
$$\phi'(\alpha) = \frac{1}{4}(\alpha+1)^2 - \frac{1}{2}(\alpha+1) - \alpha(\frac{1}{2}(\alpha+1) - 3).$$
Notice that $\phi'(\alpha)$ is a parabola that opens downward, and hence the maximum can be found by first writing ϕ' as $a\alpha^2 + b\alpha + c$, and then computing the vertex $\alpha = \frac{-b}{2a}$, which gives the global maximum for ϕ'. Performing the algebra shows that
$$\phi'(\alpha) = \frac{-1}{4}\alpha^2 + \frac{5}{2}\alpha - 1/4.$$
Hence, $\alpha^* = 5$ and $x^* = (5+1)/2 = 3$ is the primal solution, which is what we would expect when examining the graph of $x^2 - x$.

5. The Lagrangian $L(\overline{\alpha}, x, y)$ can be written as
$$L(\overline{\alpha}, x, y) = 2x^2 + y^2 - \alpha_1 x - \alpha_2 y - \alpha_3(x + 2y - 5).$$
By KKT-Condition 1,
$$\frac{\partial L(\overline{\alpha}, x, y)}{\partial x} = 4x - \alpha_1 - \alpha_3 = 0 \Rightarrow x = (\alpha_1 + \alpha_3)/4,$$
and
$$\frac{\partial L(\overline{\alpha}, x, y)}{\partial y} = 2y - \alpha_2 - 2\alpha_3 = 0 \Rightarrow y = (\alpha_2 + 2\alpha_3)/2.$$
Then substituting x for $(\alpha_1 + \alpha_3)/4$ and y for $(\alpha_2 + 2\alpha_3)/2$ gives
$$\phi'(\overline{\alpha}) = \frac{1}{8}(\alpha_1+\alpha_3)^2 + \frac{1}{4}(\alpha_2+2\alpha_3)^2 - \frac{\alpha_1}{4}(\alpha_1+\alpha_3) - \frac{\alpha_2}{2}(\alpha_2+2\alpha_3) - \alpha_3((\alpha_1+\alpha_3)/4 + \alpha_2 + 2\alpha_3 - 5).$$

6. By definition,
$$\overline{w} \cdot \overline{w} = w_1^2 + \cdots + w_n^2.$$
Then, for all $i = 1, \ldots, n$,
$$\frac{\partial(\overline{w} \cdot \overline{w})}{\partial w_i} = 2w_i.$$
Therefore,
$$\frac{\partial(\overline{w} \cdot \overline{w})}{\partial \overline{w}} = (2w_1, \ldots, 2w_n) = 2\,(w_1, \ldots, w_n) = 2\overline{w}.$$

7. Assume $(\overline{x}_1, y_1), \ldots, (\overline{x}_l, y_l)$ represent the training vectors, along with their classification labels. Then the (i, j) entry of matrix Q is $q_{ij} = y_i y_j(\overline{x} \cdot \overline{y})$, $1 \leq i, j \leq l$. Moreover, l-dimensional vector $\overline{q}$ consists of all 1's, while $(l+2)$-dimensional vector $\overline{c} = \overline{0}$. Finally, matrix X has l rows and $(l+2)$ columns. The first two columns are $(y_1, \ldots, y_l)$ and $(-y_1, \ldots, -y_l)$ respectively, while the remaining l columns form the $l \times l$ identity matrix I_l.

8. Assume $(\overline{x}_1, y_1), \ldots, (\overline{x}_6, y_6)$ represent the training vectors where

$$\overline{x}_1 = (-1,1), \overline{x}_2 = (2,2), \overline{x}_3 = (3,2), \overline{x}_4 = (-2,-4), \overline{x}_5 = (1,-1), \overline{x}_6 = (4,-2),$$

and

$$y_1 = 1, y_2 = 1, y_3 = 1, y_4 = -1, y_5 = -1, y_6 = -1.$$

Then

$$Q = \begin{pmatrix} 2 & 0 & -1 & 2 & 2 & 6 \\ 0 & 8 & 8 & 12 & 0 & -4 \\ -1 & 8 & 13 & 14 & -1 & -8 \\ 2 & 12 & 14 & 20 & 2 & 0 \\ 2 & 0 & -1 & 2 & 2 & 6 \\ 6 & -4 & -8 & 0 & 6 & 20 \end{pmatrix},$$

$\overline{q} = (1,1,1,1,1,1)$, $\overline{c} = (0,0,0,0,0,0,0,0)$, and

$$X = \begin{pmatrix} 1 & -1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 1 & 0 \\ -1 & 1 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

9.

$$\overline{w} \cdot \Phi(\overline{x}) > 1 \Leftrightarrow 1(x_1^2) + 1(x_2^2) + 0(\sqrt{2}x_1x_2) > 1 \Leftrightarrow$$
$$x_1^2 + x_2^2 > 1 \Leftrightarrow$$

$\overline{x}$ is a positive vector.

10. First suppose that

$$(\overline{u} \cdot \overline{v})^2 = (\overline{u} \cdot \overline{u})(\overline{v} \cdot \overline{v}).$$

In the proof of the Cauchy-Schwarz inequality, this would imply that

$$0 = (t\overline{u} + \overline{v}) \cdot (t\overline{u} + \overline{v}),$$

which implies

$$t\overline{u} + \overline{v} = 0,$$

or

$$\overline{v} = -t\overline{u},$$

in which case $\overline{v}$ is a scalar multiple of $\overline{u}$. Conversely, suppose $\overline{v}$ is a scalar multiple of $\overline{u}$; i.e. $\overline{v} = t\overline{u}$, for some real number t. Then

$$(\overline{u} \cdot \overline{v})^2 = ((t\overline{v}) \cdot \overline{v})^2 = (t(\overline{v} \cdot \overline{v}))^2 =$$

$$t^2(\overline{v} \cdot \overline{v})(\overline{v} \cdot \overline{v}) = ((t\overline{v}) \cdot (t\overline{v}))(\overline{v} \cdot \overline{v}) =$$
$$(\overline{u} \cdot \overline{u})(\overline{v} \cdot \overline{v}).$$

11. Let $\overline{x}_1 \in \mathcal{R}^n$ be given. Set $m = 1$ and $\theta_1 = 1$. Then by the Positive-Definite Property of k,

$$\theta_1 k(\overline{x}_1, \overline{x}_1) = 1 \cdot k(\overline{x}_1, \overline{x}_1) = k(\overline{x}_1, \overline{x}_1) \geq 0.$$

12. Certainly, the dot product is symmetric (this is one of the dot-product properties). Now for any m scalars $\theta_1, \ldots, \theta_m$ and m vectors $\overline{x}_1, \ldots, \overline{x}_m$, by the positivity property of the dot product,

$$(\theta_1 \overline{x}_1 + \cdots + \theta_n \overline{x}_n) \cdot (\theta_1 \overline{x}_1 + \cdots + \theta_n \overline{x}_n) \geq 0.$$

But after several applications of the addiviity, symmetry, and scalar-associativity properties of the dot product, we arrive at

$$(\theta_1 \overline{x}_1 + \cdots + \theta_n \overline{x}_n) \cdot (\theta_1 \overline{x}_1 + \cdots + \theta_n \overline{x}_n) = \sum_{i=1}^{m} \sum_{j=1}^{m} \theta_i \theta_j (\overline{x}_i \cdot \overline{x}_j).$$

Hence,

$$\sum_{i=1}^{m} \sum_{j=1}^{m} \theta_i \theta_j (\overline{x}_i \cdot \overline{x}_j) \geq 0,$$

and the dot-product has the property of being positive definite. Therefore, the dot product is a kernel.

13. a. Let function f represent $\Phi((1,3))$. Then

$$f(\overline{x}) = (\overline{x} \cdot (1,3))^2 = (x_1 + 3x_2)^2 = x_1^2 + 6x_1x_2 + 9x_2^2.$$

b. Let function $f(x_1, x_2)$ represent the vector

$$-2\Phi((2,1)) + 3\Phi((3,2)) - \Phi((0,1)).$$

Then

$$f(x_1, x_2) = -2(2x_1 + x_2)^2 + 3(3x_1 + 2x_2)^2 - (0x_1 + x_2)^2 =$$
$$19x_1^2 + 28x_1x_2 + 9x_2^2.$$

c. By definition,

$$\Phi((1,3)) \cdot \Phi((-2,1)) = k(,(1,3)) \cdot k(,(-2,1)) = k((-2,1),(1,3)) =$$
$$((-2,1) \cdot (1,3))^2 = (-2+3)^2 = 1.$$

d. By definition,

$$f \cdot k(,\overline{x}) = k(,(1,3)) \cdot k(,\overline{x}) = k((1,3),(x_1,x_2)) = (x_1 + 3x_2)^2 = f(x_1,x_2) = f(\overline{x}).$$

14. Given vector $\overline{x} = (x_1, \ldots, x_n) \in \mathcal{R}^n$, define $\overline{x} \otimes \overline{x}$ as the vector whose components are

$$(x_1x_1, \ldots, x_1x_n, \ldots, x_nx_1, \ldots, x_nx_n).$$

Thus $\overline{x} \otimes \overline{x} \in \mathcal{R}^{n^2}$. Now given two vectors $\overline{x}, \overline{y} \in \mathcal{R}^n$, one can readily show that

$$(\overline{x} \cdot \overline{y})^2 = (\overline{x} \otimes \overline{x}) \cdot (\overline{y} \otimes \overline{y}),$$

where the first dot product is over $\mathcal{R}^n$, and the second is the dot product over $\mathcal{R}^{n^2}$. Therefore,

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\theta_i\theta_j(\overline{x}_i\cdot\overline{x}_j)^2 =$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m}\theta_i\theta_j((\overline{x}_i\otimes\overline{x}_i)\cdot(\overline{x}_j\otimes\overline{x}_j))\geq 0,$$

since $\overline{x}_1\otimes\overline{x}_1,\ldots,\overline{x}_m\otimes\overline{x}_m$ are vectors in $\mathcal{R}^{n^2}$, and the dot product over $\mathcal{R}^{n^2}$ is a kernel, by Exercise 12.