Convex Optimization Overview (cnt'd)

Chuong B. Do

October 24, 2008

During last week's section, we began our study of ***convex optimization***, the study of mathematical optimization problems of the form,

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & x \in C. \end{array} \tag{1}$$

In a convex optimization problem, $x \in \mathbb{R}^n$ is a vector known as the ***optimization variable***, $f : \mathbb{R}^n \to \mathbb{R}$ is a ***convex function*** that we want to minimize, and $C \subseteq \mathbb{R}^n$ is a ***convex set*** describing the set of feasible solutions. From a computational perspective, convex optimization problems are interesting in the sense that any locally optimal solution will always be guaranteed to be globally optimal. Over the last several decades, general purpose methods for solving convex optimization problems have become increasingly reliable and efficient.

In these lecture notes, we continue our foray into the field of convex optimization. In particular, we explore a powerful concept in convex optimization theory known as ***Lagrange duality***. We focus on the main intuitions and mechanics of Lagrange duality; in particular, we describe the concept of the Lagrangian, its relation to primal and dual problems, and the role of the Karush-Kuhn-Tucker (KKT) conditions in providing necessary and sufficient conditions for optimality of a convex optimization problem.

1 Lagrange duality

Generally speaking, the theory of Lagrange duality is the study of optimal solutions to convex optimization problems. As we saw previously in lecture, when minimizing a differentiable convex function $f(x)$ with respect to $x \in \mathbb{R}^n$, a necessary and sufficient condition for $x^* \in \mathbb{R}^n$ to be globally optimal is that $\nabla_x f(x^*) = \mathbf{0}$. In the more general setting of convex optimization problem with constraints, however, this simple optimality condition does not work. One primary goal of duality theory is to characterize the optimal points of convex programs in a mathematically rigorous way.

In these notes, we provide a brief introduction to Lagrange duality and its applications to generic differentiable convex optimization problems of the form,

$$\begin{array}{ll} \underset{x \in \mathbb{R}^n}{\text{minimize}} & f(x) \\ \text{subject to} & g_i(x) \leq 0, \quad i = 1, \ldots, m, \\ & h_i(x) = 0, \quad i = 1, \ldots, p, \end{array} \tag{OPT}$$

where $x \in \mathbb{R}^n$ is the ***optimization variable***, $f : \mathbb{R}^n \to \mathbb{R}$ and $g_i : \mathbb{R}^n \to \mathbb{R}$ are ***differentiable convex functions***[1], and $h_i : \mathbb{R}^n \to \mathbb{R}$ are ***affine functions***.[2]

1.1 The Lagrangian

In this section, we introduce an artificial-looking construct called the "Lagrangian" which is the basis of Lagrange duality theory. Given a convex constrained minimization problem of the form (OPT), the (generalized) ***Lagrangian*** is a function $\mathcal{L} : \mathbb{R}^n \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$, defined as

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{p} \beta_i h_i(x). \tag{2}$$

Here, the first argument of the Lagrangian is a vector $x \in \mathbb{R}^n$, whose dimensionality matches that of the optimization variable in the original optimization problem; by convention, we refer to x as the ***primal variables*** of the Lagrangian. The second argument of the Lagrangian is a vector $\alpha \in \mathbb{R}^m$ with one variable α_i for each of the m convex inequality constraints in the original optimization problem. The third argument of the Lagrangian is a vector $\beta \in \mathbb{R}^p$, with one variable β_i for each of the p affine equality constraints in the original optimization problem. These elements of α and β are collectively known as the ***dual variables*** of the Lagrangian or ***Lagrange multipliers***.

Intuitively, the Lagrangian can be thought of as a modified version of the objective function to the original convex optimization problem (OPT) which accounts for each of the constraints. The Lagrange multipliers α_i and β_i can be thought of "costs" associated with violating different constraints. The key intuition behind the theory of Lagrange duality is the following:

> *For any convex optimization problem, there always exist settings of the dual variables such that the unconstrained minimum of the Lagrangian with respect to the primal variables (keeping the dual variables fixed) coincides with the solution of the original constrained minimization problem.*

We formalize this intuition when we describe the KKT conditions in Section 1.6.

---

[1]Recall that a function $f : S \to \mathbb{R}$ is convex if S is a convex set, and for any $x, y \in S$ and $\theta \in [0, 1]$, we have $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$. A function f is concave if $-f$ is convex.

[2]Recall that an affine function is a function of the form $f(x) = a^T x + b$ for some $a \in \mathbb{R}^n, b \in \mathbb{R}$. Since the Hessian of an affine function is equal to the zero matrix (i.e., it is both positive semidefinite and negative semidefinite), an affine function is both convex and concave.

1.2 Primal and dual problems

To show the relationship between the Lagrangian and the original convex optimization problem (OPT), we introduce the notions of the "primal"and "dual problems" associated with a Lagrangian:

The primal problem

Consider the optimization problem,

$$\min_x \underbrace{\left[\max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \mathcal{L}(x,\alpha,\beta)\right]}_{\text{call this } \theta_{\mathcal{P}}(x)} = \min_x \theta_{\mathcal{P}}(x). \tag{P}$$

In the equation above, the function $\theta_{\mathcal{P}} : \mathbb{R}^n \to \mathbb{R}$ is called the ***primal objective***, and the unconstrained minimization problem on the right hand side is known as the ***primal problem***. Generally, we say that a point $x \in \mathbb{R}^n$ is ***primal feasible*** if $g_i(x) \leq 0, i = 1, \ldots, m$ and $h_i(x) = 0, i = 1, \ldots, p$. We typically use the vector $x^* \in \mathbb{R}^n$ to denote the solution of (P), and we let $p^* = \theta_{\mathcal{P}}(x^*)$ denote the optimal value of the primal objective.

The dual problem

By switching the order of the minimization and maximization above, we obtain an entirely *different* optimization problem,

$$\max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \underbrace{\left[\min_x \mathcal{L}(x,\alpha,\beta)\right]}_{\text{call this } \theta_{\mathcal{D}}(x)} = \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \theta_{\mathcal{D}}(\alpha,\beta). \tag{D}$$

Here, the function $\theta_{\mathcal{D}} : \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}$ is called the ***dual objective***, and the constrained maximization problem on the right hand side is known as the ***dual problem***. Generally, we say that (α, β) are ***dual feasible*** if $\alpha_i \geq 0, i = 1, \ldots, m$. We typically use the pair of vectors $(\alpha^*, \beta^*) \in \mathbb{R}^m \times \mathbb{R}^p$ to denote the solution of (D), and we let $d^* = \theta_{\mathcal{D}}(\alpha^*, \beta^*)$ denote the optimal value of the dual objective.

1.3 Interpreting the primal problem

First, observe that the primal objective, $\theta_{\mathcal{P}}(x)$, is a convex function of x.[3] To interpret the primal problem, note that

$$\begin{aligned}
\theta_{\mathcal{P}}(x) &= \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \mathcal{L}(x,\alpha,\beta) && (4)\\
&= \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \left[f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{p} \beta_i h_i(x) \right] && (5)\\
&= f(x) + \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \left[\sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{p} \beta_i h_i(x) \right] && (6)
\end{aligned}$$

which follows from the fact that $f(x)$ does not depend on α or β. Considering only the bracketed term, notice that

- If any $g_i(x) > 0$, then maximizing the bracketed expression involves making the corresponding α_i an arbitrarily large positive number; however, if $g_i(x) \leq 0$, then the requirement that α_i be nonnegative means that the optimal setting of α_i to achieve the maximum is $\alpha_i = 0$, so that the maximum value is 0.
- Similarly, if any $h_i(x) \neq 0$, then maximizing the bracketed expression involves choosing the corresponding β_i to have the same sign as $h_i(x)$ and arbitrarily large magnitude; however, if $h_i(x) = 0$, then the maximum value is 0, independent of β_i.

Putting these two cases together, we see that if x is primal feasible (i.e., $g_i(x) \leq 0, i = 1,\ldots,m$ and $h_i(x) = 0, i = 1,\ldots,p$), then the maximum value of the bracketed expression is 0, but if any of the constraints are violated, then the maximum value is ∞. From this, we can write,

$$\theta_{\mathcal{P}}(x) = \underbrace{f(x)}_{\text{original objective}} + \underbrace{\begin{cases} 0 & \text{if } x \text{ is primal feasible} \\ \infty & \text{if } x \text{ is primal infeasible} \end{cases}}_{\text{barrier function for ``carving away'' infeasible solutions}} \quad (7)$$

Therefore, we can interpret the primal objective $\theta_{\mathcal{P}}(x)$ as a modified version of the convex objective function of the original problem (OPT), with the difference being that infeasible

[3]To see why, note that

$$\theta_{\mathcal{P}}(x) = \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \mathcal{L}(x,\alpha,\beta) = \max_{\alpha,\beta:\alpha_i\geq 0,\forall i} \left[f(x) + \sum_{i=1}^{m} \alpha_i g_i(x) + \sum_{i=1}^{p} \beta_i h_i(x) \right]. \quad (3)$$

Observe that each of the $g_i(x)$'s are convex functions in x, and since the α_i's are constrained to be nonnegative, then $\alpha_i g_i(x)$ is convex in x for each i. Similarly, each $\beta_i h_i(x)$ is convex in x (regardless of the sign of β_i) since $h_i(x)$ is linear. Since the sum of convex functions is always convex, we see that the quantity inside the brackets is a convex function of x. Finally, the maximum of a collection of convex functions is again a convex function (prove this for yourself!), so we can conclude that $\theta_{\mathcal{P}}(x)$ is a convex function of x.

solutions (i.e., x's for which some constraint is violated) have objective value ∞. Intuitively, we can consider

$$\max_{\alpha,\beta:\alpha_i\geq 0,\forall i}\left[\sum_{i=1}^{m}\alpha_i g_i(x)+\sum_{i=1}^{p}\beta_i h_i(x)\right]=\begin{cases}0 & \text{if } x \text{ is feasible for (OPT)}\\ \infty & \text{if } x \text{ is infeasible for (OPT).}\end{cases} \tag{8}$$

as a type of "barrier" function which prevents us from considering infeasible points as candidate solutions for the optimization problem.

1.4 Interpreting the dual problem

The dual objective, $\theta_{\mathcal{D}}(\alpha,\beta)$, is a concave function of α and β.[4] To interpret the dual problem, first we make the following observation:

Lemma 1. *If (α,β) are dual feasible, then $\theta_{\mathcal{D}}(\alpha,\beta)\leq p^*$*

Proof. Observe that

$$\begin{aligned}
\theta_{\mathcal{D}}(\alpha,\beta) &= \min_x \mathcal{L}(x,\alpha,\beta) && (10)\\
&\leq \mathcal{L}(x^*,\alpha,\beta) && (11)\\
&= f(x^*)+\sum_{i=1}^{m}\alpha_i g_i(x^*)+\sum_{i=1}^{p}\beta_i h_i(x^*) && (12)\\
&\leq f(x^*) = p^*. && (13)
\end{aligned}$$

Here, the first and third steps follow directly from the definitions of the dual objective function and the Lagrangian, respectively. The second step follows from the fact that the preceding expression minimized over possible values of x. The last step follows from the fact that x^* is primal feasible, (α,β) are dual feasible, and hence equation (8) implies that the latter two terms of (12) must be nonpositive. $\square$

The lemma shows that that given any dual feasible (α,β), the dual objective $\theta_{\mathcal{D}}(\alpha,\beta)$ provides a lower bound on the optimal value p^* of the primal problem. Since the dual problem involves maximizing the dual objective over the space of all dual feasible (α,β), it follows that the dual problem can be seen as a search for the tightest possible lower bound on p^*. This gives rise to a property of any primal and dual optimization problem pairs known as ***weak duality***:

[4] To see why, note that

$$\theta_{\mathcal{D}}(\alpha,\beta)=\min_x \mathcal{L}(x,\alpha,\beta)=\min_x\left[f(x)+\sum_{i=1}^{m}\alpha_i g_i(x)+\sum_{i=1}^{p}\beta_i h_i(x)\right]. \tag{9}$$

Observe that for any fixed value of x, the quantity inside the brackets is an affine function of α and β, and hence concave. Since the minimum of a collection of concave functions is also concave, we can conclude that $\theta_{\mathcal{D}}(\alpha,\beta)$ is a concave function of α and β.

Lemma 2 (Weak Duality)**.** *For any pair of primal and dual problems, $d^* \leq p^*$.*

Clearly, weak duality is a consequence of Lemma 1 using (α^*, β^*) as the dual feasible point. For some primal/dual optimization problems, an even stronger result holds, known as ***strong duality***:

Lemma 3 (Strong Duality)**.** *For any pair of primal and dual problems which satisfy certain technical conditions called* ***constraint qualifications****, then $d^* = p^*$.*

A number of different "constraint qualifications" exist, of which the most commonly invoked constraint qualification is known as ***Slater's condition***: a primal/dual problem pair satisfy Slater's condition if there exists some feasible primal solution x for which all inequality constraints are strictly satisfied (i.e., $g_i(x) < 0, i = 1, \ldots, m$). In practice, nearly all convex problems satisfy some type of constraint qualification, and hence the primal and dual problems have the same optimal value.

1.5 Complementary slackness

One particularly interesting consequence of strong duality for convex optimization problems is a property known as ***complementary slackness*** (or KKT complementarity):

Lemma 4 (Complementary Slackness)**.** *If strong duality holds, then $\alpha_i^* g(x_i^*) = 0$ for each $i = 1, \ldots, m$.*

Proof. Suppose that strong duality holds. Largely copying the proof from the last section, observe that

$$\begin{aligned}
p^* = d^* = \theta_{\mathcal{D}}(\alpha^*, \beta^*) &= \min_x \mathcal{L}(x, \alpha^*, \beta^*) && (14)\\
&\leq \mathcal{L}(x^*, \alpha^*, \beta^*) && (15)\\
&= f(x^*) + \sum_{i=1}^{m} \alpha_i^* g_i(x^*) + \sum_{i=1}^{p} \beta_i^* h_i(x^*) && (16)\\
&\leq f(x^*) = p^*. && (17)
\end{aligned}$$

Since the first and last expressions in this sequence are equal, it follows that every intermediate expression is also equal. Subtracting the left half of (17) from (16), we see that

$$\sum_{i=1}^{m} \alpha_i^* g_i(x^*) + \sum_{i=1}^{p} \beta_i^* h_i(x^*) = 0. \qquad (18)$$

Recall, however, that each α_i^* is nonnegative, each $g_i(x^*)$ is nonpositive, and each $h_i(x^*)$ is zero due to the primal and dual feasibility of x^* and (α^*, β^*), respectively. As a consequence, (18) is a summation of all nonpositive terms which equals to zero. It readily follows that all individual terms in the summation must themselves be zero (for if not, there are no compensating positive terms in the summation which would allow the overall sum to remain zero). $\square$

Complementary slackness can be written in many equivalent ways. One way, in particular, is the pair of conditions

$$\alpha_i^* > 0 \quad \Longrightarrow \quad g_i(x^*) = 0 \tag{19}$$

$$g_i(x^*) < 0 \quad \Longrightarrow \quad \alpha_i^* = 0. \tag{20}$$

In this form, we can see that whenever any α_i^* is strictly greater than zero, then this implies that the corresponding inequality constraint must hold with equality. We refer to this as an ***active constraint***. In the case of support vector machines (SVMs), active constraints are also known as ***support vectors***.

1.6 The KKT conditions

Finally, given everything so far, we can now characterize the optimal conditions for a primal dual optimization pair. We have the following theorem:

Theorem 1.1. *Suppose that $x^* \in \mathbb{R}^n$, $\alpha^* \in \mathbb{R}^m$ and $\beta^* \in \mathbb{R}^p$ satisfy the following conditions:*

1. *(Primal feasibility) $g_i(x^*) \leq 0, i = 1, \ldots, m$ and $h_i(x^*) = 0, i = 1, \ldots, p$,*
2. *(Dual feasibility) $\alpha_i^* \geq 0, i = 1, \ldots, m$,*
3. *(Complementary slackness) $\alpha_i^* g_i(x^*) = 0, i = 1, \ldots, m$, and*
4. *(Lagrangian stationarity) $\nabla_x \mathcal{L}(x^*, \alpha^*, \beta^*) = \mathbf{0}$.*

Then x^* is primal optimal and (α^*, β^*) are dual optimal. Furthermore, if strong duality holds, then any primal optimal x^* and dual optimal (α^*, β^*) must satisfy the conditions 1 through 4.

These conditions are known as the ***Karush-Kuhn-Tucker (KKT) conditions***.[5]

2 A simple duality example

As a simple application of duality, in this section, we will show how to form the dual problem for a simple convex optimization problem. Consider the convex optimization problem,

$$\begin{aligned} \underset{x \in \mathbb{R}^2}{\text{minimize}} \quad & x_1^2 + x_2 \\ \text{subject to} \quad & 2x_1 + x_2 \geq 4 \\ & x_2 \geq 1. \end{aligned}$$

[5]Incidentally, the KKT theorem has an interesting history. The result was originally derived by Karush in his 1939 master's thesis but did not catch any attention until it was rediscovered in 1950 by two mathematicians Kuhn and Tucker. A variant of essentially the same result was also derived by John in 1948. For an interesting historical account of why so many iterations of this result went unnoticed for nearly a decade, see the paper,

> Kjeldsen, T.H. (2000) A contextualized historical analysis of the Kuhn-Tucker Theorem in nonlinear programming: the impact of World War II. *Historica Mathematics* **27**: 331-361.

First, we rewrite our optimization problem in standard form as

$$\begin{aligned}
\underset{x\in\mathbb{R}^2}{\text{minimize}} \quad & x_1^2 + x_2 \\
\text{subject to} \quad & 4 - 2x_1 - x_2 \le 0 \\
& 1 - x_2 \le 0.
\end{aligned}$$

The Lagrangian is then

$$\mathcal{L}(x, \alpha) = x_1^2 + x_2 + \alpha_1(4 - 2x_1 - x_2) + \alpha_2(1 - x_2), \tag{21}$$

and the objective of the dual problem is defined to be

$$\theta_{\mathcal{D}}(\alpha) = \min_x \mathcal{L}(x, \alpha)$$

To express the dual objective in a form which depends only on α (but not x), we first observe that the the Lagrangian is differentiable in x, and in fact, is separable in the two components x_1 and x_2 (i.e., we can minimize with respect to each separately).

To minimize with respect to x_1, observe that the Lagrangian is a strictly convex quadratic function of x_1 and hence the minimum with respect to x_1 can be found by setting the derivative to zero:

$$\frac{\partial}{\partial x_1}\mathcal{L}(x, \alpha) = 2x_1 - 2\alpha_1 = 0 \quad \Longrightarrow \quad x_1 = \alpha_1. \tag{22}$$

To minimize with respect to x_2, observe that the Lagrangian is an affine function of x_2, for which the linear coefficient is precisely the derivative of the Lagrangian coefficient with respect to x_2,

$$\frac{\partial}{\partial x_2}\mathcal{L}(x, \alpha) = 1 - \alpha_1 - \alpha_2 \tag{23}$$

If the linear coefficient is non-zero, then the objective function can be made arbitrarily small by choosing the x_2 to have the opposite sign of the linear coefficient and arbitrarily large magnitude. However, if the linear coefficient is zero, then the objective function does not depend on x_2.

Putting these observations together, we have

$$\begin{aligned}
\theta_{\mathcal{D}}(\alpha) &= \min_x \mathcal{L}(x, \alpha) \\
&= \min_{x_2} \left[\alpha_1^2 + x_2 + \alpha_1(4 - 2\alpha_1 - x_2) + \alpha_2(1 - x_2)\right] \\
&= \min_{x_2} \left[-\alpha_1^2 + 4\alpha_1 + \alpha_2 + x_2(1 - \alpha_1 - \alpha_2)\right] \\
&= \begin{cases} -\alpha_1^2 + 4\alpha_1 + \alpha_2 & \text{if } 1 - \alpha_1 - \alpha_2 = 0 \\ -\infty & \text{otherwise} \end{cases}
\end{aligned}$$

so the dual problem is given by:

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^2}{\text{maximize}} \quad & \theta_{\mathcal{D}}(\alpha) \\ \text{subject to} \quad & \alpha_1 \geq 0 \\ & \alpha_2 \geq 0. \end{aligned}$$

Finally, we can simplify the dual problem by observing making the dual constraints explicit[6]:

$$\begin{aligned} \underset{\alpha \in \mathbb{R}^2}{\text{maximize}} \quad & -\alpha_1^2 + 4\alpha_1 + \alpha_2 \\ \text{subject to} \quad & \alpha_1 \geq 0 \\ & \alpha_2 \geq 0 \\ & 1 - \alpha_1 - \alpha_2 = 0. \end{aligned}$$

Notice that the dual problem is a concave quadratic program in the variables α.

3 The L_1-norm soft margin SVM

To see a more complex example of Lagrange duality in action, we derive the dual of the L_1-norm soft-margin SVM primal presented in class, as well as the corresponding KKT complementarity (i.e., complementary slackness) conditions. We have,

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i \\ \text{subject to} \quad & y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i, \quad i = 1, \ldots, m, \\ & \xi_i \geq 0, \qquad\qquad\qquad\qquad i = 1, \ldots, m. \end{aligned}$$

First, we put this into standard form, with "≤ 0" inequality constraints:

$$\begin{aligned} \underset{w,b,\xi}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i \\ \text{subject to} \quad & 1 - \xi_i - y^{(i)}(w^T x^{(i)} + b) \leq 0, \quad i = 1, \ldots, m, \\ & -\xi_i \leq 0, \qquad\qquad\qquad\qquad\quad i = 1, \ldots, m. \end{aligned}$$

Next, we form the generalized Lagrangian,[7]

$$\mathcal{L}(w, b, \xi, \alpha, \beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i + \sum_{i=1}^m \alpha_i(1 - \xi_i - y^{(i)}(w^T x^{(i)} + b)) - \sum_{i=1}^m \beta_i \xi_i,$$

[6] By this, we mean that we are moving the condition which causes $\theta_D(\alpha)$ to be $-\infty$ into the set of constraints of the dual optimization problem.

[7] Here, it is important to note that (w, b, ξ) collectively play the role of the "x" primal variables. Similarly, (α, β) collectively play the role of the "α" dual variables normally used for inequality constraints. There are no "β" dual variables here since there are no affine equality constraints in this problem.

which gives the primal and dual optimization problems:

$$\max_{\alpha,\beta:\alpha_i\geq 0,\beta_i\geq 0} \theta_{\mathcal{D}}(\alpha,\beta) \quad \text{where} \quad \theta_{\mathcal{D}}(\alpha,\beta) := \min_{w,b,\xi} \mathcal{L}(w,b,\xi,\alpha,\beta), \tag{SVM-D}$$

$$\min_{w,b,\xi} \theta_{\mathcal{P}}(w,b,\xi) \quad \text{where} \quad \theta_{\mathcal{P}}(w,b,\xi) := \max_{\alpha,\beta:\alpha_i\geq 0,\beta_i\geq 0} \mathcal{L}(w,b,\xi,\alpha,\beta). \tag{SVM-P}$$

To get the dual problem in the form shown in the lecture notes, however, we still have a little more work to do. In particular,

1. **Eliminating the primal variables.** To eliminate the primal variables from the dual problem, we compute $\theta_{\mathcal{D}}(\alpha,\beta)$ by noticing that

   $$\theta_{\mathcal{D}}(\alpha,\beta) = \min\nolimits_{w,b,\xi} \quad \mathcal{L}(w,b,\xi,\alpha,\beta)$$

   is an unconstrained optimization problem, where the objective function $\mathcal{L}(w,b,\xi,\alpha,\beta)$ is differentiable. The Lagrangian is a strictly convex quadratic function of w, so for any fixed (α,β), if $(\hat{w},\hat{b},\hat{\xi})$ minimize the Lagrangian, it must be the case that

   $$\nabla_w \mathcal{L}(\hat{w},\hat{b},\hat{\xi},\alpha,\beta) = \hat{w} - \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} = 0. \tag{24}$$

   Furthermore, the Lagrangian is linear in b and ξ; by reasoning analogous to that described in the simple duality example from the previous section, we can set the derivatives with respect to b and ξ to zero, and add the resulting conditions as explicit constraints in the dual optimization problem:

   $$\frac{\partial}{\partial b}\mathcal{L}(\hat{w},\hat{b},\hat{\xi},\alpha,\beta) = -\sum_{i=1}^m \alpha_i y^{(i)} = 0 \tag{25}$$

   $$\frac{\partial}{\partial \xi_i}\mathcal{L}(\hat{w},\hat{b},\hat{\xi},\alpha,\beta) = C - \alpha_i - \beta_i = 0. \tag{26}$$

   We can use these conditions to compute the dual objective as

   $$\begin{aligned}
\theta_{\mathcal{D}}(\alpha,\beta) &= \mathcal{L}(\hat{w},\hat{b},\hat{\xi}) \\
&= \frac{1}{2}\|\hat{w}\|^2 + C\sum_{i=1}^m \hat{\xi}_i + \sum_{i=1}^m \alpha_i(1 - \hat{\xi}_i - y^{(i)}(\hat{w}^T x^{(i)} + \hat{b})) - \sum_{i=1}^m \beta_i\hat{\xi}_i \\
&= \frac{1}{2}\|\hat{w}\|^2 + C\sum_{i=1}^m \hat{\xi}_i + \sum_{i=1}^m \alpha_i(1 - \hat{\xi}_i - y^{(i)}(\hat{w}^T x^{(i)})) - \sum_{i=1}^m \beta_i\hat{\xi}_i \\
&= \frac{1}{2}\|\hat{w}\|^2 + \sum_{i=1}^m \alpha_i(1 - y^{(i)}(\hat{w}^T x^{(i)})),
\end{aligned}$$

   where the first equality follows from the optimality of $(\hat{w},\hat{b},\hat{\xi})$ for fixed (α,β), the second equality uses the definition of the generalized Lagrangian, and the third and fourth equalities follow from (25) and (26), respectively. Finally, to use (24), observe that

$$
\begin{aligned}
\frac{1}{2}\|\hat{w}\|^2 + \sum_{i=1}^m \alpha_i(1 - y^{(i)}(\hat{w}^T x^{(i)})) &= \sum_{i=1}^m \alpha_i + \frac{1}{2}\|\hat{w}\|^2 - \hat{w}^T \sum_{i=1}^m \alpha_i y^{(i)} x^{(i)} \\
&= \sum_{i=1}^m \alpha_i + \frac{1}{2}\|\hat{w}\|^2 - \|\hat{w}\|^2 \\
&= \sum_{i=1}^m \alpha_i - \frac{1}{2}\|\hat{w}\|^2 \\
&= \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_i y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle.
\end{aligned}
$$

Therefore, our dual problem (with no more primal variables and all constraints made explicit) is simply

$$
\begin{aligned}
\underset{\alpha,\beta}{\text{maximize}} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_i y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle \\
\text{subject to} \quad & \alpha_i \geq 0, & i = 1, \ldots, m, \\
& \beta_i \geq 0, & i = 1, \ldots, m, \\
& \alpha_i + \beta_i = C, & i = 1, \ldots, m, \\
& \sum_{i=1}^m \alpha_i y^{(i)} = 0.
\end{aligned}
$$

2. **KKT complementary.** KKT complementarity requires that for any primal optimal (w^*, b^*, ξ^*) and dual optimal (α^*, β^*),

$$
\begin{aligned}
\alpha_i^*(1 - \xi_i^* - y^{(i)}(w^{*T} x^{(i)} + b^*)) &= 0 \\
\beta_i^* \xi_i^* &= 0
\end{aligned}
$$

for $i = 1, \ldots, m$. From the first condition, we see that if $\alpha_i^* > 0$, then in order for the product to be zero, then $1 - \xi_i^* - y^{(i)}(w^{*T} x^{(i)} + b^*) = 0$. It follows that

$$y^{(i)}(w^{*T} x^{(i)} + b^*) \leq 1$$

since $\xi^* \geq 0$ by primal feasibility. Similarly, if $\beta_i^* > 0$, then $\xi_i^* = 0$ to ensure complementarity. From the primal constraint, $y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i$, it follows that

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

Finally, since $\beta_i^* > 0$ is equivalent to $\alpha_i^* < C$ (since $\alpha^* + \beta_i^* = C$), we can summarize the KKT conditions as follows:

$$
\begin{aligned}
\alpha_i^* < C \quad &\Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \geq 1, \\
\alpha_i^* > 0 \quad &\Rightarrow \quad y^{(i)}(w^{*T} x^{(i)} + b^*) \leq 1.
\end{aligned}
$$

or equivalently,

$$
\begin{aligned}
\alpha_i^* = 0 \quad &\Rightarrow \quad y^{(i)}(w^{*T}x^{(i)} + b^*) \geq 1, \\
0 < \alpha_i^* < C \quad &\Rightarrow \quad y^{(i)}(w^{*T}x^{(i)} + b^*) = 1, \\
\alpha_i^* = C \quad &\Rightarrow \quad y^{(i)}(w^{*T}x^{(i)} + b^*) \leq 1.
\end{aligned}
$$

3. **Simplification.** We can tidy up our dual problem slightly by observing that each pair of constraints of the form

$$\beta_i \geq 0 \qquad\qquad\qquad \alpha_i + \beta_i = C$$

is equivalent to the single constraint, $\alpha_i \leq C$; that is, if we solve the optimization problem

$$
\begin{aligned}
\underset{\alpha,\beta}{\text{maximize}} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_i y^{(i)} y^{(j)} \langle x^{(i)}, x^{(j)} \rangle \\
\text{subject to} \quad & 0 \leq \alpha_i \leq C, \qquad\qquad\qquad i = 1, \ldots, m, \\
& \sum_{i=1}^m \alpha_i y^{(i)} = 0.
\end{aligned}
\tag{27}
$$

and subsequently set $\beta_i = C - \alpha_i$, then it follows that (α, β) will be optimal for the previous dual problem above. This last form, indeed, is the form of the soft-margin SVM dual given in the lecture notes.

4 Directions for further exploration

In many real-world tasks, 90% of the challenge involves figuring out how to write an optimization problem in a convex form. Once the correct form has been found, a number of pre-existing software packages for convex optimization have been well-tuned to handle different specific types of optimization problems. The following constitute a small sample of the available tools:

- commerical packages: CPLEX, MOSEK
- MATLAB-based: CVX, Optimization Toolbox (linprog, quadprog), SeDuMi
- libraries: CVXOPT (Python), GLPK (C), COIN-OR (C)
- SVMs: LIBSVM, SVM-light
- machine learning: Weka (Java)

In particular, we specifically point out CVX as an easy-to-use generic tool for solving convex optimization problems easily using MATLAB, and CVXOPT as a powerful Python-based library which runs independently of MATLAB.[8] If you're interested in looking at some of the other packages listed above, they are easy to find with a web search. In short, if you need a specific convex optimization algorithm, pre-existing software packages provide a rapid way to prototype your idea without having to deal with the numerical trickiness of implementing your own complete convex optimization routines.

Also, if you find this material fascinating, make sure to check out Stephen Boyd's class, EE364: Convex Optimization I, which will be offered during the Winter Quarter. The textbook for the class (listed as [1] in the References) has a wealth of information about convex optimization and is available for browsing online.

References

[1] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge UP, 2004. Online: `http://www.stanford.edu/~boyd/cvxbook/`

[8] CVX is available at `http://www.stanford.edu/~boyd/cvx/` and CVXOPT is available at `http://www.ee.ucla.edu/~vandenbe/cvxopt/`.