Lagrange Multipliers Tutorial in the Context of Support Vector Machines

Baxter Tyson Smith, B.Sc., B.Eng., Ph.D. Candidate

Faculty of Engineering and Applied Science
Memorial University of Newfoundland
St. John's, Newfoundland, Canada
baxter@engr.mun.ca

June 3, 2004

Contents

1 Introduction

The purpose of this tutorial is to explain how lagrange multipliers work in the context of Support Vector Machines (SVMs). During my research on SVMs, I have read many papers and tutorials that talk about SVMs in detail, but when they get to the part about solving the constrained optimization equations they just say …*and this can be solved easily using Lagrange Multipliers*…. This tutorial aims to answer the questions that the others don't - at least the questions that I had when learning about SVMs. If you have any other questions that should go here, please let me know.

I'll begin with a very simple tutorial on lagrange multipliers. I'll tell what they are and why one would use them. I'll also give a few examples on using them with equality contraints. Then, I'll give an explanation on how to use them with inequality contraints, since this is how they are used in the context of SVMs. I'll give examples here also - one solving the Primal Lagrangian and one solving the Dual Lagrangian. The last section contains a list of Frequently Asked Questions that I had on Lagrange Multipliers. If you have any others, please email me and I'll add them.

2 Lagrange Multipliers

Lagrange Multipliers are a mathematical method used to solve constrained optimization problems of differentiable functions. What does this mean? Well, basically you have some function $f(x_1, \ldots, x_n) : R^n \rightarrow R$ that you want to optimize (ie find the min or max extremes). Hold on, if it were this simple, you could just use the *second derivative test*. Well, in addition to this function, you also have a constraint $g(x_1, \ldots, x_n) = 0$. So, we are trying to optimize f, while constraining f with g. You can think of a constraint as a boundary. As a laymans terms example, say we are in a room and we want to find out the highest distance that we can throw a ball. Well, we are constrained by the ceiling!!! We can't throw the ball any higher than that.

At the heart of Lagrange Multipliers is the following equation:

$$\nabla f(x) = \lambda \nabla g(x) \tag{1}$$

This says that the gradient of f is equal to some multiplier (lagrange multiplier) times the gradient of g. How this equation came about is explained in Section 6. Also, remember the form of g:

$$g(x) = 0 \tag{2}$$

Often, and especially in the context of SVMs, equations 1 and 2 are combined into one equation called the *Lagrangian*:

$$L(x, \lambda) = f(x) - \lambda g(x) \tag{3}$$

Using this equation, we look for points where:

$$\nabla L(x, \lambda) = 0 \tag{4}$$

That is essentially it. I'll give an example or two to illustrate.

2.1 Example 1: One Equality Constraint

Problem: Given,

$$\begin{aligned} f(x,y) &= 2 - x^2 - 2y^2 & (5) \\ g(x,y) &= x + y - 1 = 0 & (6) \end{aligned}$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x,y,\lambda) &= f(x,y) - \lambda g(x,y) \qquad (7) \\ &= 2 - x^2 - 2y^2 - \lambda(x+y-1) \qquad (8) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda \nabla g(x,y) = 0 \qquad (9)$$

which gives us:

$$\begin{aligned} \frac{\partial}{\partial x} L(x,y,\lambda) &= -2x - \lambda = 0 \qquad (10) \\ \frac{\partial}{\partial y} L(x,y,\lambda) &= -4y - \lambda = 0 \qquad (11) \\ \frac{\partial}{\partial \lambda} L(x,y,\lambda) &= x + y - 1 = 0 \qquad (12) \end{aligned}$$

From Equation 10 and 11, we have $x = 2y$. Substituting this into Equation 12 gives $x = \frac{2}{3}$, $y = \frac{1}{3}$. These values give $\lambda = -\frac{4}{3}$ and $f = \frac{4}{3}$.

2.2 Example 2: One Equality Constraint

Problem: Given,

$$\begin{aligned} f(x,y) &= x + 2y \qquad (13) \\ g(x,y) &= y^2 + xy - 1 = 0 \qquad (14) \end{aligned}$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x,y,\lambda) &= f(x,y) - \lambda g(x,y) \qquad (15) \\ &= x + 2y - \lambda(y^2 + xy - 1) \qquad (16) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda \nabla g(x,y) = 0 \qquad (17)$$

which gives us:

$$\begin{aligned}
\frac{\partial}{\partial x}L(x,y,\lambda) &= 1-\lambda y = 0 \qquad (18)\\
\frac{\partial}{\partial y}L(x,y,\lambda) &= 2-2\lambda y - \lambda x = 0 \qquad (19)\\
\frac{\partial}{\partial \lambda}L(x,y,\lambda) &= y^2 + xy - 1 = 0 \qquad (20)
\end{aligned}$$

This gives $x = 0$, $y = \pm 1$, $\lambda = \pm 1$ and $f = \pm 2$.

3 Multiple Constraints

Lagrange Multipliers works just as well with multiple constraints. In essence, we are just adding another boundary to the problem. Keep in mind that with equality constraints, we are not staying within a boundary, we are actually touching the boundary. We'll get to staying within a boundary in Section 4. A simple re-wording of the Lagrangian takes into account multiple constraints:

$$L(x, \lambda) = f(x) - \sum_i \lambda_i g_i(x) \tag{21}$$

Here $g_i(x)$ and λ_i are the multiple constraints (denoted by i), and associated Lagrange Multipliers. Note that each constraint has its own multiplier. Again, we look for points where:

$$\nabla L(x, \lambda) = 0 \tag{22}$$

It is not much different to solve than the single constraint case. Here is an example to illustrate.

3.1 Example 3: Two Equality Constraints

Problem: Given,

$$\begin{aligned} f(x, y) &= x^2 + y^2 && (23) \\ g_1(x, y) &= x + 1 = 0 && (24) \\ g_2(x, y) &= y + 1 = 0 && (25) \end{aligned}$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x, y, \lambda) &= f(x, y) - \lambda_1 g_1(x, y) - \lambda_2 g_2(x, y) && (26) \\ &= x^2 + y^2 - \lambda_1(x + 1) - \lambda_2(y + 1) && (27) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 22):

$$\nabla L(x, y, \lambda) = \nabla f(x, y) - \lambda_1 \nabla g_1(x, y) - \lambda_2 \nabla g_2(x, y) = 0 \tag{28}$$

which gives us:

$$\begin{aligned}
\frac{\partial}{\partial x}L(x,y,\lambda) &= 2x-\lambda_1=0 \quad (29)\\
\frac{\partial}{\partial y}L(x,y,\lambda) &= 2y-\lambda_2=0 \quad (30)\\
\frac{\partial}{\partial \lambda_1}L(x,y,\lambda) &= x+1=0 \quad (31)\\
\frac{\partial}{\partial \lambda_2}L(x,y,\lambda) &= y+1=0 \quad (32)
\end{aligned}$$

Equation 31 gives $x=-1$. Equation 32 gives $y=-1$. Substituting this into Equation 29 and 30 gives $\lambda_1=-2$, $\lambda_2=-2$ and $f=2$.

4 Inequality Constraints

Now we are getting closer to the Lagrange Multipliers representation of SVMs. This section details using Lagrange Multipliers with Inequality Constraints (ie $g(x) \leq 0$,$g(x) \geq 0$). For these types of problems, the formulation of the Lagrangian remains the same as in Equation 3. The constraints are handled by the Lagrange Multipliers themselves. The following equations details the rules on how the Lagrange Multipliers encode the inequality constraints:

$$
\begin{aligned}
g(x) \geq 0 \quad &\Rightarrow \quad \lambda \geq 0 && (33)\\
g(x) \leq 0 \quad &\Rightarrow \quad \lambda \leq 0 && (34)\\
g(x) = 0 \quad &\Rightarrow \quad \lambda \text{ is unconstrained} && (35)
\end{aligned}
$$

So, handling inequality constraints isn't any harder than handling equality constraints. All we have to do is restrict the values of the Lagrange Multipliers accordingly. Time for an example.

4.1 Example 4: One Inequality Constraint

Problem: Given,

$$
\begin{aligned}
f(x,y) &= x^3 + y^2 && (36)\\
g(x,y) &= x^2 - 1 \geq 0 && (37)
\end{aligned}
$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$
\begin{aligned}
L(x,y,\lambda) &= f(x,y) - \lambda g(x,y) && (38)\\
&= x^3 + y^2 - \lambda(x^2 - 1) && (39)
\end{aligned}
$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda \nabla g(x,y) = 0 \qquad (40)$$

which gives us:

$$\frac{\partial}{\partial x} L(x,y,\lambda) = 3x^2 - 2\lambda x = 0 \qquad (41)$$

$$
\begin{aligned}
\frac{\partial}{\partial y} L(x, y, \lambda) &= 2y = 0 && (42) \\
\frac{\partial}{\partial \lambda} L(x, y, \lambda) &= x^2 - 1 = 0 && (43)
\end{aligned}
$$

Furthermore, we require that:

$$\lambda \geq 0 \tag{44}$$

since we are dealing with an inequality constraint.

From Equation 42, we have $y = 0$. From Equation 43, we have $x = \pm 1$. Substituting this into Equation 41 gives $\lambda = \pm\frac{3}{2}$. Since we require that $\lambda \geq 0$, then $\lambda = \frac{3}{2}$. This gives $x = 1$, $y = 0$ and $f = 1$.

4.2 Example 5: Two Inequality Constraints

Problem: Given,

$$
\begin{aligned}
f(x, y) &= x^3 + y^3 && (45) \\
g_1(x, y) &= x^2 - 1 \geq 0 && (46) \\
g_2(x, y) &= y^2 - 1 \geq 0 && (47)
\end{aligned}
$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$
\begin{aligned}
L(x, y, \lambda) &= f(x, y) - \lambda_1 g_1(x, y) - \lambda_2 g_2(x, y) && (48) \\
&= x^3 + y^3 - \lambda_1(x^2 - 1) - \lambda_2(y^2 - 1) && (49)
\end{aligned}
$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x, y, \lambda) = \nabla f(x, y) - \lambda_1 \nabla g_1(x, y) - \lambda_2 \nabla g_2(x, y) = 0 \tag{50}$$

which gives us:

$$
\begin{aligned}
\frac{\partial}{\partial x} L(x, y, \lambda) &= 3x^2 - 2\lambda_1 x = 0 && (51) \\
\frac{\partial}{\partial y} L(x, y, \lambda) &= 3y^2 - 2\lambda_2 y = 0 && (52) \\
\frac{\partial}{\partial \lambda_1} L(x, y, \lambda) &= x^2 - 1 = 0 && (53) \\
\frac{\partial}{\partial \lambda_2} L(x, y, \lambda) &= y^2 - 1 = 0 && (54)
\end{aligned}
$$

Furthermore, we require that:

$$\lambda_1 \geq 0 \tag{55}$$

$$\lambda_2 \geq 0 \tag{56}$$

since we are dealing with a inequality constraints.

From Equations 53 and 54, we have $x = \pm 1$ and $y = \pm 1$. Substituting $x = \pm 1$ into Equation 51 gives $\lambda_1 = \pm\frac{3}{2}$. Since we require that $\lambda_1 \geq 0$, then $\lambda_1 = \frac{3}{2}$ is the only valid choice for λ_1. Likewise, substituting $y = \pm 1$ into Equation 52 gives $\lambda_2 = \pm\frac{3}{2}$. Since we require that $\lambda_2 \geq 0$, then $\lambda_2 = \frac{3}{2}$. This gives $x = 1$, $y = 1$ and $f = 2$.

4.3 Example 6: Two Inequality Constraints

Problem: Given,

$$\begin{aligned} f(x,y) &= x^3 + y^3 & (57)\\ g_1(x,y) &= x^2 - 1 \geq 0 & (58)\\ g_2(x,y) &= y^2 - 1 \leq 0 & (59) \end{aligned}$$

Find the extreme values.

Solution: First, we put the equations into the form of a Lagrangian:

$$\begin{aligned} L(x,y,\lambda) &= f(x,y) - \lambda_1 g_1(x,y) - \lambda_2 g_2(x,y) & (60)\\ &= x^3 + y^3 - \lambda_1(x^2 - 1) - \lambda_2(y^2 - 1) & (61) \end{aligned}$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$\nabla L(x,y,\lambda) = \nabla f(x,y) - \lambda_1 \nabla g_1(x,y) - \lambda_2 \nabla g_2(x,y) = 0 \tag{62}$$

which gives us:

$$\begin{aligned} \frac{\partial}{\partial x} L(x,y,\lambda) &= 3x^2 - 2\lambda_1 x = 0 & (63)\\ \frac{\partial}{\partial y} L(x,y,\lambda) &= 3y^2 - 2\lambda_2 y = 0 & (64)\\ \frac{\partial}{\partial \lambda_1} L(x,y,\lambda) &= x^2 - 1 = 0 & (65)\\ \frac{\partial}{\partial \lambda_2} L(x,y,\lambda) &= y^2 - 1 = 0 & (66) \end{aligned}$$

Furthermore, we require that:

$$\lambda_1 \geq 0 \tag{67}$$

$$\lambda_2 \leq 0 \tag{68}$$

since we are dealing with a inequality constraints.

From Equations 65 and 66, we have $x = \pm 1$ and $y = \pm 1$. Substituting $x = \pm 1$ into Equation 63 gives $\lambda_1 = \pm\frac{3}{2}$. Since we require that $\lambda_1 \geq 0$, then $\lambda_1 = \frac{3}{2}$ is the only valid choice for λ_1. Likewise, substituting $y = \pm 1$ into Equation 64 gives $\lambda_2 = \pm\frac{3}{2}$. Since we require that $\lambda_2 \leq 0$, then $\lambda_2 = -\frac{3}{2}$. This gives $x = 1$, $y = -1$ and $f = 0$.

5 Application to SVMs

This section will go over detailed examples of how Lagrange Multipliers work with SVMs. Applying Lagrange Multipliers to SVMs is exactly the same as we did above. Before we jump into an example, I want to present the Karush-Kuhn-Tucker (KKT) conditions. These conditions must also be satisfied when performing any constraint-based optimization. I didn't mention it before, because we didn't need it to solve the simple equations, but you can verify for yourself that they apply to the above examples as well.

5.1 Karush-Kuhn-Tucker Conditions

There are five KKT conditions that affect all of our constraint based optimizations. I won't go into a proof, I'll just present them. They are:

$$
\begin{aligned}
\frac{\partial}{\partial \overline{w}} L(\overline{w}, b, \lambda) = \overline{w} - \sum_i \lambda_i y_i \overline{x}_i &= 0 \qquad (69)\\
\frac{\partial}{\partial b} L(\overline{w}, b, \lambda) = -\sum_i \lambda_i y_i &= 0 \qquad (70)\\
y_i \left[\langle \overline{w}, \overline{x} \rangle + b\right] - 1 &\geq 0 \qquad (71)\\
\lambda_i &\geq 0 \qquad (72)\\
\lambda_i (y_i \left[\langle \overline{w}, \overline{x} \rangle + b\right] - 1) &= 0 \qquad (73)
\end{aligned}
$$

So, anytime we apply a constraint-based optimization, we must ensure that these conditions are satisfied.

5.2 Example 7: Applying Lagrange Multipliers Directly to SVMs

Problem: Lets assume that we have two classes of two-dimensional data to separate. Lets also assume that each class consists of only one point. These points are:

$$
\begin{aligned}
\overline{x}_1 &= A_1 = (1, 1)\\
\overline{x}_2 &= B_1 = (2, 2) \qquad (74)
\end{aligned}
$$

Find the hyperplane that separates these two classes.

Solution: From SVM theory, we know that the equations are:

$$
\begin{aligned}
f(\overline{w}) &= \frac{1}{2}\left\|\overline{w}\right\|^2 && (75)\\
g_i(\overline{w}, b) &= y_i\left[\langle\overline{w}, \overline{x}_i\rangle + b\right] - 1 \geq 0 && (76)
\end{aligned}
$$

A common question here is *Why isn't g_i a function of $\overline{x}_i$?*. The answer of course is that x_i isn't a variable - each x_i has a value which we know from Equation 74. We can expand $g_i(\overline{w}, b)$ a bit further:

$$
\begin{aligned}
g_1(\overline{w}, b) &= \left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1 \geq 0 && (77)\\
g_2(\overline{w}, b) &= -\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] - 1 \geq 0 && (78)
\end{aligned}
$$

Next, we put the equations into the form of a Lagrangian:

$$
\begin{aligned}
L(\overline{w}, b, \lambda) &= f(\overline{w}) - \lambda_1 g_1(\overline{w}, b) - \lambda_2 g_2(\overline{w}, b)\\
&= \frac{1}{2}\left\|\overline{w}\right\|^2 - \lambda_1(\left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1) - \lambda_2(-\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] - 1)\\
&= \frac{1}{2}\left\|\overline{w}\right\|^2 - \lambda_1(\left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1) + \lambda_2(\left[\langle\overline{w}, \overline{x}_2\rangle + b\right] + 1) && (79)
\end{aligned}
$$

and we solve for the gradient of the Lagrangian (Equation 4):

$$
\nabla L(\overline{w}, b, \lambda) = \nabla f(\overline{w}) - \lambda_1 \nabla g_1(\overline{w}, b) - \lambda_2 \nabla g_2(\overline{w}, b) = 0 \qquad (80)
$$

which gives us:

$$
\begin{aligned}
\frac{\partial}{\partial\overline{w}} L(\overline{w}, b, \lambda) &= \overline{w} - \lambda_1\overline{x}_1 + \lambda_2\overline{x}_2 = 0 && (81)\\
\frac{\partial}{\partial b} L(\overline{w}, b, \lambda) &= -\lambda_1 + \lambda_2 = 0 && (82)\\
\frac{\partial}{\partial\lambda_1} L(\overline{w}, b, \lambda) &= \left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1 = 0 && (83)\\
\frac{\partial}{\partial\lambda_2} L(\overline{w}, b, \lambda) &= \left[\langle\overline{w}, \overline{x}_2\rangle + b\right] + 1 = 0 && (84)
\end{aligned}
$$

This gives us enough equations to solve this analytically. Equating Equations 83 and 84 we get:

$$
\begin{aligned}
\left[\langle\overline{w}, \overline{x}_1\rangle + b\right] - 1 &= \left[\langle\overline{w}, \overline{x}_2\rangle + b\right] + 1 = 0 && (85)\\
\langle\overline{w}, \overline{x}_1\rangle + b - 1 &= \langle\overline{w}, \overline{x}_2\rangle + b + 1 && (86)\\
\langle\overline{w}, \overline{x}_1\rangle - 1 &= \langle\overline{w}, \overline{x}_2\rangle + 1 && (87)\\
\langle\overline{w}, \overline{x}_1\rangle - \langle\overline{w}, \overline{x}_2\rangle &= 2 && (88)\\
\langle\overline{w}, [\overline{x}_1 - \overline{x}_2]\rangle &= 2 && (89)
\end{aligned}
$$

We have the values of $\overline{x}_1$ and $\overline{x}_2$ from Equation 74. This leaves $\overline{w}$ as the unknown. We can break $\overline{w}$ down into its components:

$$\overline{w} = (w_1, w_2) \tag{90}$$

Adding these into the mix we get:

$$\begin{aligned}
\langle \overline{w}, [\overline{x}_1 - \overline{x}_2] \rangle &= 2 && (91)\\
\langle (w_1, w_2), [(1,1) - (2,2)] \rangle &= 2 && (92)\\
\langle (w_1, w_2), (-1,-1) \rangle &= 2 && (93)\\
-w_1 - w_2 &= 2 && (94)\\
w_1 &= -(2 + w_2) && (95)
\end{aligned}$$

Adding values to Equation 81 and combining with Equation 82 gives us:

$$\begin{aligned}
(w_1, w_2) - \lambda_1(1,1) + \lambda_2(2,2) &= 0 && (96)\\
(w_1, w_2) - \lambda_1(1,1) + \lambda_1(2,2) &= 0 && (97)\\
(w_1, w_2) + \lambda_1(1,1) &= 0 && (98)
\end{aligned}$$

Which yields:

$$w_1 + \lambda_1 = 0 \tag{99}$$

and

$$w_2 + \lambda_1 = 0 \tag{100}$$

Equating these we get:

$$w_1 = w_2 \tag{101}$$

Putting this result back into Equation 95 gives:

$$w_1 = w_2 = -1 \tag{102}$$

Using this in either of Equations 99 or 100 will give:

$$\lambda_1 = \lambda_2 = 1 \tag{103}$$

And finally, using this in Equations 83 and 84 give:

$$\begin{aligned}
b &= 1 - \langle \overline{w}, \overline{x}_1 \rangle && (104)\\
&= -1 - \langle \overline{w}, \overline{x}_2 \rangle && (105)\\
&= 1 - \langle (-1,-1), (1,1) \rangle && (106)\\
&= -1 - \langle (-1,-1), (2,2) \rangle && (107)\\
&= 3 && (108)
\end{aligned}$$

Note that this result also satisfies all of the KKT conditions including:

$$\lambda_i(y_i\,[\langle \overline{w}, \overline{x}_i\rangle + b] - 1) = 0 \tag{109}$$

ie:

$$\begin{aligned}
\lambda_1([\langle \overline{w}, \overline{x}_1\rangle + b] - 1) &= 0 & (110)\\
\lambda_2([\langle \overline{w}, \overline{x}_2\rangle + b] + 1) &= 0 & (111)\\
([\langle(-1,-1),(1,1)\rangle + 3] - 1) &= -2 + 3 - 1 = 0 & (112)\\
([\langle(-1,-1),(2,2)\rangle + 3] + 1) &= -4 + 3 + 1 = 0 & (113)
\end{aligned}$$

and the inequality constaints:

$$\begin{aligned}
\lambda_1 &\geq 0 & (114)\\
\lambda_2 &\geq 0 & (115)
\end{aligned}$$

This two point problem seems overly complicated. In general, it is. Anything greater than a few points cannot be solved analytically. Usually, the SVM optimization problem can be solved analytically only when the number of training data is very small, or for the separable case when it is known beforehand which of the training data become support vectors. In most real-world problems, this must be solved numerically.

5.3 Example 8: Using the Wolfe Dual to Apply Lagrange Multipliers to SVMs

Problem: Lets assume that we have two classes of two-dimensional data to separate. Lets also assume that each class consists of only one point. These points are:

$$\begin{aligned}
\overline{x}_1 &= A_1 = (1,1)\\
\overline{x}_2 &= B_1 = (2,2) & (116)
\end{aligned}$$

Find the hyperplane that separates these two classes.

Solution: Note that this is the same problem as the previous one, but we are going to solve it in a different way. This time we are going to use the Wolfe dual of the Lagrangian to to it. It is supposed to make things simpler. Lets see if it does. The equation for the primal representation of the Lagrangian for a SVM is:

$$\begin{aligned} L(\overline{w}, b, \lambda) &= \frac{1}{2}\|w\|^2 - \lambda_1([\langle \overline{w}, \overline{x}_1\rangle + b] - 1) + \lambda_2([\langle \overline{w}, \overline{x}_2\rangle + b] + 1) && (117) \\ &= \frac{1}{2}\|w\|^2 - \lambda_1(\langle \overline{w}, \overline{x}_1\rangle + b - 1) + \lambda_2(\langle \overline{w}, \overline{x}_2\rangle + b + 1) && (118) \\ &= \frac{1}{2}\|w\|^2 - \lambda_1\langle \overline{w}, \overline{x}_1\rangle - \lambda_1 b + \lambda_1 + \lambda_2\langle \overline{w}, \overline{x}_2\rangle + \lambda_2 b + \lambda_2) && (119) \\ &= \frac{1}{2}\|w\|^2 - \lambda_1\langle \overline{w}, \overline{x}_1\rangle + \lambda_2\langle \overline{w}, \overline{x}_2\rangle - \lambda_1 b + \lambda_2 b + \lambda_1 + \lambda_2 && (120) \end{aligned}$$

If we substitute Equations 81 and 82 into this formulation we get:

$$\begin{aligned} L(\lambda) \;=\; & \frac{1}{2}\|\lambda_1\overline{x}_1 - \lambda_2\overline{x}_2\|^2 \\ & - \lambda_1\langle \lambda_1\overline{x}_1 - \lambda_2\overline{x}_2, \overline{x}_1\rangle + \lambda_2\langle \lambda_1\overline{x}_1 - \lambda_2\overline{x}_2, \overline{x}_2\rangle && (121) \\ & + b(\lambda_2 - \lambda_1) \\ & + \lambda_1 + \lambda_2 \\ =\; & \frac{1}{2}(\lambda_1^2\langle \overline{x}_1, \overline{x}_1\rangle - 2\lambda_1\lambda_2\langle \overline{x}_1, \overline{x}_2\rangle + \lambda_2^2\langle \overline{x}_2, \overline{x}_2\rangle) \\ & - \lambda_1^2\langle \overline{x}_1, \overline{x}_1\rangle + \lambda_1\lambda_2\langle \overline{x}_1, \overline{x}_2\rangle + \lambda_1\lambda_2\langle \overline{x}_1, \overline{x}_2\rangle - \lambda_2^2\langle \overline{x}_2, \overline{x}_2\rangle && (122) \\ & + b(0) \\ & + \lambda_1 + \lambda_2 \\ =\; & \lambda_1 + \lambda_2 + \lambda_1\lambda_2\langle \overline{x}_1, \overline{x}_2\rangle - \frac{1}{2}\lambda_1^2\langle \overline{x}_1, \overline{x}_1\rangle - \frac{1}{2}\lambda_2^2\langle \overline{x}_2, \overline{x}_2\rangle && (123) \end{aligned}$$

which is the equation for the Wolfe Dual Lagrangian. Keep in mind that this is also subject to $\lambda_i \geq 0$ and all the KKT constraints [69-73]. Equation 82, which is a KKT constraint, can be rewritten as:

$$-\sum_i \lambda_i y_i = 0 \qquad (124)$$

and since it is a constraint, we must also take it into account when taking the gradient of the Lagrangian. We add it the same way we add any constraint. Thus the Dual Lagrangian becomes:

$$\begin{aligned} L(\lambda, \gamma) \;=\; & \lambda_1 + \lambda_2 + \lambda_1\lambda_2\langle \overline{x}_1, \overline{x}_2\rangle - \frac{1}{2}\lambda_1^2\langle \overline{x}_1, \overline{x}_1\rangle - \frac{1}{2}\lambda_2^2\langle \overline{x}_2, \overline{x}_2\rangle \\ & - \gamma(\lambda_1 - \lambda_2) && (125) \end{aligned}$$

$$\begin{aligned} &= \lambda_1 + \lambda_2 + \lambda_1\lambda_2 \langle \overline{x}_1, \overline{x}_2 \rangle - \frac{1}{2}\lambda_1^2 \langle \overline{x}_1, \overline{x}_1 \rangle - \frac{1}{2}\lambda_2^2 \langle \overline{x}_2, \overline{x}_2 \rangle \\ &\quad - \gamma\lambda_1 + \gamma\lambda_2 \qquad (126) \end{aligned}$$

Now we just need to find the Lagrange Multipliers, λ_1 and λ_2. To do this we solve for the gradient of the Dual Lagrangian which gives us:

$$\frac{\partial}{\partial\lambda_1}L(\lambda, \gamma) = 1 + \lambda_2 \langle \overline{x}_1, \overline{x}_2 \rangle - \lambda_1 \langle \overline{x}_1, \overline{x}_1 \rangle - \gamma = 0 \qquad (127)$$

$$\frac{\partial}{\partial\lambda_2}L(\lambda, \gamma) = 1 + \lambda_1 \langle \overline{x}_1, \overline{x}_2 \rangle - \lambda_2 \langle \overline{x}_2, \overline{x}_2 \rangle + \gamma = 0 \qquad (128)$$

$$\frac{\partial}{\partial\gamma}L(\lambda, \gamma) = -\lambda_1 + \lambda_2 = 0 \qquad (129)$$

Solving this gives $\lambda_1 = \lambda_2 = 1$ and $\gamma = 3$. It is interesting to note that we don't need to solve for γ explicitly. That is, we don't need it to solve for $\overline{w}$ or b. We can use KKT conditions 69 and 73 to solve for these.

6 Why do Lagrange Multipliers Work?

This is actually quite interesting. When we have a function, f, constrained by another function, g, we have an extremum (max or min) when the normals to these functions are parallel, that is, the functions are tangent to each other. This gives the fundamental equation for Lagrange Multipliers:

$$\nabla f(x) = \lambda \nabla g(x) \tag{130}$$

where λ is the lagrange multiplier. The equation says that the gradients are parallel, but may be of different sizes or different directions ie λ is a scaling factor.

7 FAQ

What are SVMs? SVMs stands for Support Vector Machines.

Where does the name Support Vector Machines come from? Support Vector Machines are a type of *learning machine* that uses *kernels* to extend linear discriminant machines into the nonlinear domain. As such they can be used to discriminate, or tell the difference, between two classes of data. In laymans terms, they draw a line or plane between the two sets of data and whatever is on one side is of class A and whatever is on the other side is on class B.

What are Lagrange Multipliers? The Lagrange Multipliers are a scaling factor by which the gradient of the function, f that you want to find the extremem of is equal to the gradient of the constraints, g_i.

Where does the name Lagrange Multipliers come from? Some dude named Lagrange. Did you think it was from some dude named Multiplier? If so, please let me know.

When would I use Lagrange Multipliers? You can use Lagrange Multipliers whenever you have a function, f, that is constrained by a function, g, or functions, g_i, and you want to find the extremum (largest or smallest value) of that function.

Are there methods to use for constrained optimization other than Lagrange Multipliers? Yes. For instance, the method of *Parametrizing the Constraint Set*. Its explanation is beyond the scope of this paper, but you can look it up.

Why use Lagrange Multipliers instead of other methods for constrained optimization? The main reason for using Lagrange Multipliers is that once you know how they work it is very easy to setup the problem. The result, is a fairly complicated system of equations, but there are methods to solve these. Plus, if you use a computer then they are fairly simple to solve. A method like parametrizing the constraint set is harder to get started, because it can be hard to find a parametrization of the given constraint set.

My constraint isn't in the correct form. It looks like: $g(x_1, \ldots, x_n) = c$, where c is a constant. How do I put it in the correct form? Just move the constant, c, to the left: $g(x_1, \ldots, x_n) - c = 0$.

How do I get from the Primal to the Dual form of the SVM Lagrangian? There are two equations that we use to get from the Primal to the Dual form of the Lagrangian. They are Equations 69 and 70 from the KKT conditions. We just substitute these into the Primal Equation and rearrange. Also, note that the constraints must also be applied to the Dual form using additional Lagrange Multipliers. See Example 8.

Solving Example 8 using the Dual form of the SVM Lagrangian didn't seem any easier than just using the Primal form. So, why do we bother to use it? It doesn't seem any easier for this simple problem when solving analytically, but it does make solving easier when using more complicated situations and when solving numerically.

What is the *Second Derivative Test*? This is outside the scope of this tutorial. It is helpful to know it though. Maybe I'll give a tutorial on that later.

What material did you derive this tutorial from? There are several other tutorials and papers that I derived this tutorial from. Dan Klein's *Lagrange Multipliers without Permanent Scarring*, and more...

I have a question about Lagrange Multipliers in the Context of SVMs that you didn't cover here. Where do I find the answer? Email me your question.