Introduction to the Karush-Kuhn-Tucker (KKT) Conditions

Illinois Institute of Technology
Department of Applied Mathematics

Adam Rumpf
arumpf@hawk.iit.edu

April 20, 2018

1 Lagrangian Multipliers

We preface our discussion of the KKT conditions with a simpler class of problem since it leads to a simpler analysis. Lagrangian relaxation is a technique that applies to optimization problems subject to equality constraints. In general, any optimization problem with n variables $\mathbf{x} = (x_1, \ldots, x_n)$ and m constraints can be written in the form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) \\ \text{s.t.} \quad & g_i(\mathbf{x}) = 0 \qquad i = 1, \ldots, m \end{aligned}$$

We will assume for the remainder of the discussion that any functions involved in the problem are differentiable, but note that subgradient versions of most of these results exist. The above program is asking us to find the vector $\mathbf{x}$ that minimizes the value of function f, but restricted only to the set of $\mathbf{x}$ such that $g_i(\mathbf{x}) = 0$ holds for all i.

The Lagrangian relaxation of this problem is

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x})$$

subject to no constraints. This involves m new variables λ_i, called the Lagrange multipliers (in the case of linear programming, these are the dual variables). We can think of this as eliminating the constraints, but adding a penalty cost to the objective. If any of the terms $g_i(\mathbf{x})$ is nonzero, we incur a unit penalty cost of λ_i. If the penalty costs are chosen carefully, the solution to the relaxed problem will be exactly the same as the solution to the original problem.

1.1 Derivation

We will derive the method of Lagrange multipliers for the two-dimensional problem

$$\begin{aligned} \min \quad & f(x, y) \\ \text{s.t.} \quad & g(x, y) = 0 \end{aligned}$$

We can imagine the level sets of $f(x, y)$ on the plane, and $g(x, y)$ as a parametric curve in the plane. Our goal is to find the point (x, y) on the curve h that reaches the lowest possible level set of f. Imagine walking along g and watching how the value of f changes. For f to reach a minimum, there must be some point on this walk where its value momentarily does not change. This could occur for one of two reasons: either because g is tangent to a contour of f, or because f, itself, flattens out.

To check whether g is tangent to a contour of f, remember that the gradient $\nabla_{x,y} = (\frac{\partial}{\partial x}, \frac{\partial}{\partial y})$ of a function is a vector orthogonal to its contour lines. If g is tangent to a contour of f, then the gradient of g should be a scalar multiple of the gradient of f. That is, there should be some constant λ such that

$$\nabla_{x,y} f = \lambda \nabla_{x,y} g$$

This condition is also true if f flattens out, since in that case $\nabla_{x,y} f = \mathbf{0}$ and we can just use $\lambda = 0$. We have just found a necessary condition for some point (x, y) to be an optimal solution to the original problem: it must satisfy $\nabla_{x,y} f(x, y) = \lambda \nabla_{x,y} g(x, y)$ for some λ (stationary), and it must satisfy $g(x, y) = 0$ (feasible).

We can incorporate all of this into a single equation. We define the Lagrangian as

$$\mathcal{L}(x, y, \lambda) = f(x, y) + \lambda g(x, y)$$

and solve $\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) = \mathbf{0}$. This encapsulates both our stationarity and our feasibility requirement. To see this, expand

$$\begin{aligned}
\nabla_{x,y,\lambda} \mathcal{L}(x, y, \lambda) &= \nabla_{x,y,\lambda} [f(x, y) + \lambda g(x, y)] \\
&= \nabla_{x,y,\lambda} f(x, y) + \nabla_{x,y,\lambda} \lambda g(x, y) \\
&= \begin{bmatrix} \frac{\partial}{\partial x} f(x, y) \\ \frac{\partial}{\partial y} f(x, y) \\ \frac{\partial}{\partial \lambda} f(x, y) \end{bmatrix} + \begin{bmatrix} \frac{\partial}{\partial x} \lambda g(x, y) \\ \frac{\partial}{\partial y} \lambda g(x, y) \\ \frac{\partial}{\partial \lambda} \lambda g(x, y) \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial}{\partial x} f(x, y) \\ \frac{\partial}{\partial y} f(x, y) \\ 0 \end{bmatrix} + \begin{bmatrix} \lambda \frac{\partial}{\partial x} g(x, y) \\ \lambda \frac{\partial}{\partial y} g(x, y) \\ g(x, y) \end{bmatrix}
\end{aligned}$$

Looking at only the first two coordinates gives

$$\begin{bmatrix} \frac{\partial}{\partial x} f(x, y) \\ \frac{\partial}{\partial y} f(x, y) \end{bmatrix} + \begin{bmatrix} \lambda \frac{\partial}{\partial x} g(x, y) \\ \lambda \frac{\partial}{\partial y} g(x, y) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial x} f(x, y) \\ \frac{\partial}{\partial y} f(x, y) \end{bmatrix} + \lambda \begin{bmatrix} \frac{\partial}{\partial x} g(x, y) \\ \frac{\partial}{\partial y} g(x, y) \end{bmatrix} = \nabla_{x,y} f(x, y) + \lambda \nabla_{x,y} g(x, y)$$

while the third coordinate is simply $g(x, y)$. Setting the entire expression equal to $\mathbf{0}$ then requires that $g(x, y) = 0$, and that $\nabla_{x,y} f(x, y) + \lambda \nabla_{x,y} g(x, y) = 0$ for some λ, which occurs if and only if $\nabla_{x,y} f(x, y) = \lambda \nabla_{x,y} g(x, y)$ for some λ.

1.2 General Result

This result generalizes to any number of variables and constraints. Given an optimization problem

$$\begin{aligned}
\min \quad & f(x_1, \ldots, x_n) \\
\text{s.t.} \quad & g_i(x_1, \ldots, x_n) = 0 \qquad i = 1, \ldots, m
\end{aligned}$$

the Lagrangian is defined as

$$\mathcal{L}(x_1, \ldots, x_n, \lambda_1, \ldots, \lambda_m) = f(x_1, \ldots, x_n) + \sum_{i=1}^{m} \lambda_i g_i(x_1, \ldots, x_n)$$

and our necessary condition for optimality is that

$$\nabla_{x_1,\ldots,x_n,\lambda_1,\ldots,\lambda_m}\mathcal{L}(x_1,\ldots,x_n,\lambda_1,\ldots,\lambda_m)=\mathbf{0}$$

Stated another way, given

$$\begin{aligned}\min\quad & f(\mathbf{x})\\ \text{s.t.}\quad & g_i(\mathbf{x})=0 \qquad i=1,\ldots,m\end{aligned}$$

the following two conditions are necessary for $\mathbf{x}$ to be an optimal solution:

- **Stationarity:** There exists some set of multipliers $\boldsymbol{\lambda}$ such that

$$\nabla_{\mathbf{x}}f(\mathbf{x})+\sum_{i=1}^{m}\lambda_i\nabla_{\mathbf{x}}g_i(\mathbf{x})=\mathbf{0}$$

- **Feasibility:** $$g_i(\mathbf{x})=0 \qquad i=1,\ldots,m$$

Note that this represents a system of $n+m$ equations and $n+m$ unknowns (the n variables plus the m multipliers), so solving it gives us not only a locally optimal solution $\mathbf{x}$, but also their corresponding multipliers $\boldsymbol{\lambda}$. Specifically, the system we need to solve is $\nabla_{\mathbf{x},\boldsymbol{\lambda}}\mathcal{L}(\mathbf{x},\boldsymbol{\lambda})=\mathbf{0}$.

1.3 Example

Consider the problem

$$\begin{aligned}\min\quad & x+y\\ \text{s.t.}\quad & x^2+y^2=1\end{aligned}$$

In this example, $f(x,y)=x+y$ and $g(x,y)=x^2+y^2-1$. The Lagrangian is

$$\mathcal{L}(x,y,\lambda)=x+y+\lambda(x^2+y^2-1)$$

Applying the method of Lagrange multipliers, we should solve $\nabla\mathcal{L}(x,y,\lambda)=\mathbf{0}$. This gives us the system

$$\nabla\mathcal{L}(x,y,\lambda)=\begin{bmatrix}1+2x\lambda\\ 1+2y\lambda\\ x^2+y^2-1\end{bmatrix}$$

Setting this equal to $\mathbf{0}$ and solving gives the solution $(x,y,\lambda)=(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2})$, so the optimal solution is $(x,y)=(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})$, which has an objective value of $f(-\frac{\sqrt{2}}{2},-\frac{\sqrt{2}}{2})=-\sqrt{2}$.

2 KKT Conditions

The Karush-Kuhn-Tucker (KKT) conditions are a generalization of Lagrange multipliers, and give a set of necessary conditions for optimality for systems involving both equality and inequality constraints. Suppose we have a problem with variables $\mathbf{x} = (x_1, \ldots, x_n)$, m equality constraints, and ℓ inequality constraints. Such a problem has the general form

$$\begin{aligned} \min_{\mathbf{x}} \quad & f(\mathbf{x}) & \\ \text{s.t.} \quad & g_i(\mathbf{x}) = 0 & i = 1, \ldots, m \\ & h_j(\mathbf{x}) \leq 0 & j = 1, \ldots, \ell \end{aligned}$$

Again, we form a relaxed problem by eliminating the constraints and adding a penalty cost for each constraint. This gives

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \mu_j h_j(\mathbf{x})$$

We now have a multiplier λ_i for each equality constraint and μ_j for each inequality constraint. These are called KKT multipliers. If there are only equality constraints, then we are left with only the Lagrange multipliers (and all the same results) from before.

Much of the same reasoning from the method of Lagrange multipliers still applies here. We can define the objective above as the Lagrangian $\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$, and we find that necessary conditions for optimality include that $\nabla_{\mathbf{x}} \mathcal{L} = \mathbf{0}$ (stability) and $\nabla_{\boldsymbol{\lambda}} \mathcal{L} = \mathbf{0}$ (feasibility of the inequality constraints). However, it is not necessary to also have $\nabla_{\boldsymbol{\mu}} \mathcal{L} = \mathbf{0}$, since this is equivalent to enforcing $h_j(\mathbf{x}) = 0$, which gives us only part of the feasible set. As a result, we cannot simply say that $\nabla_{\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}} \mathcal{L} = \mathbf{0}$. This leaves us with only $n + m$ equations, but $n + m + \ell$ variables, so we need some additional information to determine the values of the multipliers $\boldsymbol{\mu}$.

Note that, in the previous section, we often switched between the forms $\nabla f = \lambda \nabla g$ and $\nabla f + \lambda \nabla g = \mathbf{0}$, since the lack of constraints for λ made them equivalent. We will find that in this case the sign of μ actually matters, so we will need to start being more careful about the signs of the terms.

2.1 Derivation

We start by considering the simplified two-dimensional system

$$\begin{aligned} \min \quad & f(x, y) \\ \text{s.t.} \quad & h(x, y) \leq 0 \end{aligned}$$

which involves only inequality constraints. Again, we can imagine the level sets of f in the plane. The constraint $h(x, y) \leq 0$ defines a region within the plane. There are two possibilities for the optimal solution: either it lies completely within the feasible region, in which case $h(x, y) < 0$ and f is at an extremum, or it lies on the boundary, in which case $h(x,y)=0$ and the boundary of $h(x,y)$ is tangent to a contour of f.

If the optimum occurs where $h(x,y)<0$, then the inequality constraint has no effect on the problem, and can be ignored. Otherwise, $h(x,y)=0$. In either case, the constraint $\mu h(x,y)=0$ must be satisfied for some μ. Either $h(x,y)=0$, in which case the value of μ does not matter, or $h(x,y)<0$, in which case we must have $\mu=0$. This is complementary slackness.

We do need to consider the sign of μ, however. Our goal is to be able to use this variable as part of the objective of the relaxed problem

$$\min \quad f(x,y)+\mu h(x,y)$$

As before, it is necessary for optimality that this expression be stationary at (x,y), which is true if we have $\nabla_{x,y} f(x,y)+\mu\nabla_{x,y} h(x,y)=\mathbf{0}$. We can rearrange this to find that

$$\mu=-\frac{[\nabla f(x,y)]_k}{[\nabla h(x,y)]_k} \qquad k=1,2$$

Recall that the gradient of a function is the direction of steepest ascent. Since we are dealing with the case in which the minimum of f lies outside the region defined by h, the gradient of f should point into the region. Otherwise, a local minimum would occur inside the region or on the other side of it, and in either case (x,y) would not be the minimum of the feasible set. As for h, since the boundary of the region is defined by $h(x,y)=0$ and the interior is defined by $h(x,y)<0$, h must decrease as we move towards the interior, meaning that the gradient of h should point out of the region. These two directions are opposite, so $[\nabla f(x,y)]_k$ and $[\nabla h(x,y)]_k$ must have opposite signs for $k=1,2$, meaning that $-\frac{[\nabla f(x,y)]_k}{[\nabla h(x,y)]_k}$, and thus μ, is always positive.

Because of this, we need to be careful when we write the stationary condition for maximization instead of minimization. If, instead, we were attempting to maximize f, its gradient would point towards the outside of the region defined by h. Then $\nabla f(x,y)$ and $\nabla h(x,y)$ would have the same direction, which would force μ to be negative. We could include this as an alternate constraint, but by convention we still require μ to be positive, and simply change the sign of the penalty term in the Lagrangian, resulting in $\nabla_{x,y} f(x,y)-\mu\nabla_{x,y} h(x,y)=\mathbf{0}$.

2.2 General Result

Given a general constrained nonlinear optimization problem

$$\begin{aligned} \min \quad & f(\mathbf{x}) & \\ \text{s.t.} \quad & g(\mathbf{x})=0 & \qquad i=1,\dots,m \\ & h(\mathbf{x})\le 0 & \qquad j=1,\dots,\ell \end{aligned}$$

or the maximization version, the KKT conditions are a set of necessary conditions that any optimal solution $\mathbf{x}=(x_1,\dots,x_n)$ must satisfy. Specifically, there must exist multipliers $\boldsymbol{\lambda}=(\lambda_1,\dots,\lambda_m)$ and $\boldsymbol{\mu}=(\mu_1,\dots,\mu_\ell)$ such that the following hold.

- **Stationarity:** If minimizing,

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} g(\mathbf{x}) + \sum_{j=1}^{\ell} \mu_j \nabla_{\mathbf{x}} h(\mathbf{x}) = \mathbf{0}$$

  or if maximizing,

$$\nabla_{\mathbf{x}} f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i \nabla_{\mathbf{x}} g(\mathbf{x}) - \sum_{j=1}^{\ell} \mu_j \nabla_{\mathbf{x}} h(\mathbf{x}) = \mathbf{0}$$

- **Primal Feasibility:**

$$g_i(\mathbf{x}) = 0 \qquad i = 1, \dots, m$$
$$h_j(\mathbf{x}) \leq 0 \qquad j = 1, \dots, \ell$$

- **Dual Feasibility:** $\mu_j \geq 0 \qquad j = 1, \dots, \ell$

- **Complementary Slackness:** $\mu_j h_j(\mathbf{x}) = 0 \qquad j = 1, \dots, \ell$

This does not work for all programs, meaning that there are certain optimization problems where, for a given optimal solution $\mathbf{x}$, there do not exist any multipliers $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ that satisfy the KKT conditions. In order for a solution to exist, f, g_i, and h_j must satisfy certain regularity conditions. There are a few broad classes of constraint that are known to always satisfy these conditions.

Importantly, if all functions g_i and h_j are affine, we automatically have regularity. Another important class of functions are those satisfying Slater's condition, which requires that the program be convex, and that there exist some $\mathbf{x}$ satisfying $g_i(\mathbf{x}) = 0$ for $i = 1, \dots, m$ and $h_j(\mathbf{x}) < 0$ for $j = 1, \dots, \ell$ (i.e. there must be some feasible solution where all inequality constraints are inactive). It is also important to note that, for a convex program satisfying the regularity conditions with continuously differentiable constraints, the KKT conditions are both necessary and sufficient for the global optimum.

2.3 Duality

The expressions used in the relaxed objective is called the Lagrangian, just like from before. The Lagrangian is

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i g_i(\mathbf{x}) + \sum_{j=1}^{\ell} \mu_j h_j(\mathbf{x})$$

We can define the Lagrangian dual function as

$$z(\boldsymbol{\lambda}, \boldsymbol{\mu}) = \min_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu})$$

Then the dual problem is

$$\begin{aligned} \max_{\boldsymbol{\lambda},\boldsymbol{\mu}} \quad & z(\boldsymbol{\lambda},\boldsymbol{\mu}) \\ \text{s.t.} \quad & \mu_j \geq 0 \qquad j = 1,\ldots,\ell \end{aligned}$$

The dual is always convex, regardless of the primal problem. We always have weak duality, meaning that any feasible solution to the dual is smaller than any feasible solution to the primal problem. If the primal problem satisfies Slater's condition, then we also have strong duality. In particular, this is always the case for linear programs.

References

[1] D.P. Bertsekas. *Nonlinear Programming.* 2nd ed. Athena Scientific, Belmont, MA, 1999.