Nonlinear Programming, Part I: Optimality Conditions Under Differentiability Assumptions

John MacLaren Walsh, Ph.D.

1 Reference

- **Nonlinear Programming, 2nd Ed.**, Dimitri P. Bertsekas. Athena Scientific, 1999.

2 Purpose

This lecture primarily presented results concerning the optimization of differentiable functions via conditions involving first and second order derivatives and Lagrange multipliers.

3 Unconstrained Optimization (Bertsekas 1.1)

Throughout this section, consider a twice continuously differentiable function $f : \mathbb{R}^N \to \mathbb{R}$. If $f(\mathbf{x}^*) \leq f(\mathbf{x})$ for every $\mathbf{x} \in \mathbb{R}^N$, then we say that f has a *global minimum* at $\mathbf{x}^*$.

We say that f has a *local minimum* at $\mathbf{x}^*$ if there is some ball $\mathcal{B}_\epsilon(\mathbf{x}^*) := \{\mathbf{x} \mid \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ surrounding $\mathbf{x}^*$ such that for any $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}^*)$, $f(\mathbf{x}^*) \leq f(\mathbf{x})$.

Convex functions are special in that any local minimum of a convex function must be a global minimum, and further that the set of global minima of a convex function must be convex.

To see this, proceed with a proof by contradiction by supposing not. Then there would exist $\mathbf{x}_1$ a local minimum, and another location $\mathbf{x}_2$ with lower cost, i.e. with $f(\mathbf{x}_1) > f(\mathbf{x}_2)$. Because $\mathbf{x}_1$ is a local minimum, there must exist a ball $\mathcal{B}_\epsilon(\mathbf{x}_1)$ surrounding $\mathbf{x}_1$ such that $f(\mathbf{x}') \geq f(\mathbf{x}_1)$ for all $\mathbf{x}' \in \mathcal{B}_\epsilon(\mathbf{x}_1)$, on the other hand consider the line segment between $\mathbf{x}_1$ and $\mathbf{x}_2$, parameterized as $\mathbf{x}(\lambda) = (1-\lambda)\mathbf{x}_1 + \lambda\mathbf{x}_2$ with $\lambda \in (0, 1)$. Take λ small enough that $\mathbf{x}(\lambda) \in \mathcal{B}_\epsilon(\mathbf{x}_1)$, because $\mathbf{x}_1$ was supposed to be a local minimum $f(\mathbf{x}(\lambda)) \geq f(\mathbf{x}_1)$. However, because f is convex $f(\mathbf{x}(\lambda)) \leq (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_2) < (1-\lambda)f(\mathbf{x}_1) + \lambda f(\mathbf{x}_1) = f(\mathbf{x}_1)$. Hence there is a contradiction.

Additionally, strictly convex functions can have at most one global minimum.

3.1 Necessary Derivative Conditions for Local Optimality

A necessary condition for f to have a local minimum at $\mathbf{x}^*$ is for the gradient (vector of partial derivatives) to be equal to zero, i.e. for

$$\nabla f(\mathbf{x}^*) = \left[\frac{\partial f}{\partial x_1}(\mathbf{x}^*), \frac{\partial f}{\partial x_2}(\mathbf{x}^*), \ldots, \frac{\partial f}{\partial x_N}(\mathbf{x}^*) \right]^T = \mathbf{0} \tag{1}$$

and for the hessian matrix to be positive semi-definite, i.e.

$$\nabla^2 f(\mathbf{x}^*) = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_N} \\ \frac{\partial^2 f}{\partial x_1 \partial x_2} & \frac{\partial^2 f}{\partial x_2^2} & \cdots & \frac{\partial^2 f}{\partial x_2 \partial x_N} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_1 \partial x_N} & \cdots & \cdots & \frac{\partial^2 f}{\partial x_N^2} \end{bmatrix} \quad \text{positive semi definite} \tag{2}$$

(Recall that a real symmetric matrix $\mathbf{A}$ is positive semi-definite if for any $\mathbf{x} \in \mathbf{R}^N$, the quadratic form $\mathbf{x}^T\mathbf{A}\mathbf{x} \geq 0$. Equivalently, a matrix is positive semi-definite if and only if all of its eigenvalues are ≥ 0.)

To see why these are necessary conditions, begin with the first order condition and proceed with a proof by contradiction. By contradiction assumption, there exists a local minimum $\mathbf{x}^*$ with gradient $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$. Because $\mathbf{x}^*$ is a local minimum, there must exist a ball $\mathcal{B}_\epsilon(\mathbf{x}^*)$ surrounding $\mathbf{x}^*$ such that $f(\mathbf{x}') \geq f(\mathbf{x}^*)$ for all $\mathbf{x}' \in \mathcal{B}_\epsilon(\mathbf{x}^*)$. Proceeding with a first order Taylor series approximation of f in the neighborhood of $\mathbf{x}^*$ we have

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\nabla f(\mathbf{x}^*))^T(\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|) \tag{3}$$

Consider then $\mathbf{x} = \mathbf{x}^* - \epsilon' \nabla f(\mathbf{x}^*)$ with ϵ' chosen small enough so that $\mathbf{x}$ is in $\mathcal{B}_\epsilon(\mathbf{x}^*)$, at this $\mathbf{x}$ the Taylor series gives

$$\begin{aligned} f(\mathbf{x}) &= f(\mathbf{x}^*) + (\nabla f(\mathbf{x}^*))^T(-\epsilon' \nabla f(\mathbf{x}^*)) + o(\epsilon') \\ &= f(\mathbf{x}^*) - \epsilon' \|\nabla f(\mathbf{x})\|_2^2 + o(\epsilon) \\ &< f(\mathbf{x}^*) \quad \text{for small enough } \epsilon'. \end{aligned}$$

which is a contradiction to the local optimality of $f(\mathbf{x}^*)$. Hence we see that if $\mathbf{x}^*$ is to be a local optimum of f, the gradient $\nabla f(\mathbf{x}^*) = \mathbf{0}$. Moving on next to the Hessian matrix, we gain proceed with a proof by contradiction. By contradiction assumption $\nabla^2 f(\mathbf{x}^*)$ is not positive semi-definite, so that there exists a direction $\mathbf{y}$ such that $\mathbf{y}^T(\nabla^2 f(\mathbf{x}^*))\mathbf{y} < 0$. We take the Taylor series approximation of f out to second order and obtain

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\nabla f(\mathbf{x}^*))^T(\mathbf{x} - \mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|^2) \tag{4}$$

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^T \nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|^2) \tag{5}$$

Selecting $\mathbf{x} = \mathbf{x}^* + \epsilon_1 \mathbf{y}$ with ϵ_1 small enough for $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}^*)$, we have

$$f(\mathbf{x}) = f(\mathbf{x}^*) + \epsilon_1^2 \mathbf{y}^T \nabla^2 f(\mathbf{x}^*)\mathbf{y} + o(\epsilon_1) \tag{6}$$

$$< f(\mathbf{x}^*) \quad \text{for small enough } \epsilon_1 \tag{7}$$

contradicting the local optimality of $\mathbf{x}^*$. Hence we have shown that a local minimum of a twice continuously differentiable function f must have gradient zero and Hessian positive semi-definite.

Points $\mathbf{x}^*$ for which $\nabla f(\mathbf{x}^*) = \mathbf{0}$ are called stationary points. A point $\mathbf{x}^*$ being a stationary point and having a positive semi-definite Hessian is a necessary, but not sufficient condition for $\mathbf{x}^*$ to be a local minimum.

A sufficient, but not necessary, condition for $\mathbf{x}^*$ to be a local minimum is for $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and for $\nabla^2 f(\mathbf{x}^*)$ to be positive definite. Indeed, again by the Taylor series, this implies that

$$f(\mathbf{x}) = f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)\nabla^2 f(\mathbf{x}^*)(\mathbf{x} - \mathbf{x}^*) + o(\|\mathbf{x} - \mathbf{x}^*\|^2) \tag{8}$$

which shows that (due to positive definiteness) there is a neighborhood $\mathcal{B}_\epsilon(\mathbf{x}^*)$ for which any $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}^*)$, $f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

We noted that there were equivalent versions of these theorems for local maxima. Namely, that a necessary (but not sufficient) condition for $\mathbf{x}^*$ to be a local maximum of f is that it be a stationary point (i.e. $\nabla f(\mathbf{x}^*) = \mathbf{0}$) and that the Hessian $\nabla^2 f(\mathbf{x})$ is negative semi-definite. Similarly, a sufficient (but not necessary) condition for $\mathbf{x}^*$ to be a local maximum of f is that it be a stationary point $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and the Hessian is negative definite $\nabla^2 f(\mathbf{x})$.

When the Hessian matrix has some eigenvalues which are positive and some eigenvalues which are negative at a stationary points $\mathbf{x}^*$, that stationary point is said to be a *saddle point*.

It was also interesting to note that when the Hessian matrix has zero eigenvalues the associated stationary point is called degenerate and some very wild things can occur. Two examples on the homework this week provide examples of some of the things that can happen. There is a branch of theory (called catastrophe theory in nonlinear dynamical systems – where the surface involved is the potential function of an associated dynamical system) which classifies such degenerate stationary points.

4 Constrained Optimization via Lagrange Multiplier Theory (Bertsekas ch. 3)

4.1 Equality Constraints

We next considered the solutions to to optimization problems of the form

$$\min_{\mathbf{x} | h_1(\mathbf{x}) = 0, h_2(\mathbf{x}) = 0, \ldots, h_I(\mathbf{x}) = 0} f(\mathbf{x}) \tag{9}$$

where f is again a twice continuously differentiable function taking $\mathbb{R}^N \to \mathbb{R}$, as are also the constraint functions $h_1, \ldots, h_I$.

We say that $\mathbf{x}^*$ is a local minimum of such an optimization problem if obeys the constraints, i.e., $h_i(\mathbf{x}^*) = 0$ and if there is a ball surrounding $\mathbf{x}^*$, $\mathcal{B}_\epsilon(\mathbf{x}^*) := \{\mathbf{x} \,|\, \|\mathbf{x} - \mathbf{x}^*\| < \epsilon\}$ (with $\epsilon > 0$), such that for any $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}^*)$ and such that $h_i(\mathbf{x}) = 0,\ \forall i \in \{1, \ldots, I\}$, $f(\mathbf{x}) > f(\mathbf{x}^*)$.

In studying conditions for local optimality for optimization problems in the class, it is useful to define a function

$$L(\mathbf{x}, \boldsymbol{\lambda}) = f(\mathbf{x}) + \sum_{i=1}^{I} \lambda_i h_i(\mathbf{x}) \tag{10}$$

called the *Lagrangian*, where the new auxiliary variables $\lambda_1, \ldots, \lambda_I$ are called Lagrange multipliers.

A point $\mathbf{x}^*$ is said to be *regular* if the gradients of the constraints $\{\nabla h_1(\mathbf{x}^*), \ldots, \nabla h_I(\mathbf{x}^*)\}$ are linearly independent.

A necessary condition for a regular $\mathbf{x}^*$ to be a local minimum is that

$$\nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*) = \mathbf{0} \tag{11}$$

Here the gradient can be taken with respect to both $\mathbf{x}$ and $\boldsymbol{\lambda}$ because the gradient with respect to $\boldsymbol{\lambda}$ when set to zero is just the collection of constraints for the problem.

To see why this is a necessary condition, begin by observing that the existence of such a collection of Langrange multipliers is equivalent to requiring that $\nabla f(\mathbf{x}^*)$ be expressible as a linear combination of $\{\nabla h_1(\mathbf{x}^*), \nabla h_2(\mathbf{x}^*), \ldots, \nabla h_I(\mathbf{x}^*)\}$, i.e.

$$\nabla f(\mathbf{x}^*) = -\sum_{i=1}^{I} \lambda_i^* \nabla h_i(\mathbf{x}^*) \tag{12}$$

Such a linear combination must obviously be possible (due to simple linear algebra) if $I = N$, since then regularity would dictate that $\{\nabla h_1(\mathbf{x}^*), \ldots, \nabla h_I(\mathbf{x}^*)\}$ span the entire vector space. Hence, restrict attention next to the more interesting case when $I < N$. Proceed again with a proof by contradiction. If no such Lagrange multipliers existed, then there would exist a direction $\mathbf{y}$ such that $(\nabla f(\mathbf{x}^*))^T \mathbf{y} < 0$ but for which $[\nabla h_1(\mathbf{x}^*), \ldots, h_I(\mathbf{x}^*)]^T \mathbf{y} = 0$. But then we would be able to move along this direction away from $\mathbf{x}^*$ using a differentiable curve $\mathbf{x}(t)$ such that $\mathbf{x}(0) = \mathbf{x}^*$ with $\frac{d}{dt}\mathbf{x}(0) = \mathbf{y}$ and for which $h_i(\mathbf{x}(t)) = 0$. (This is possible due to the implicit function theorem, the full rank derivative sub-matrix requirements for which are the root of the regularity assumption.) However, $f(\mathbf{x}(t))$ would then be decreasing in the neighborhood of $\mathbf{0}$, since by the chain rule $\frac{d}{dt} f(\mathbf{x}(t))|_{t=0} = (\nabla f(\mathbf{x}^*))^T \mathbf{y} < 0$, contradicting the local optimality of $\mathbf{x}^*$.

An additional necessary condition for $\mathbf{x}^*$ to be a regular local minimum is that

$$\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}, \boldsymbol{\lambda}^*) \mathbf{y} \geq 0 \tag{13}$$

for any $\mathbf{y}$ such that $(\nabla h_i(\mathbf{x}^*))^T \mathbf{y} = 0$ for all $i \in \{1, \ldots, I\}$.

To see this, observe that the statement is again vacuous if $I = N$, for then the only $\mathbf{y}$ satisfying the condition $\mathbf{y}$ such that $(\nabla h_i(\mathbf{x}^*))^T \mathbf{y} = 0$ for all $i \in \{1, \ldots, I\}$ is $\mathbf{y} = \mathbf{0}$, hence focus on the more interesting case that $I < N$. Again proceed with a proof by contradiction. Then there would be a $\mathbf{y}$ such that $(\nabla h_i(\mathbf{x}^*))^T \mathbf{y} = 0$ for all $i \in \{1, \ldots, I\}$, but for which $\mathbf{y}^T \nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*) \mathbf{y} < 0$ for all $i \in \{1, \ldots, I\}$. Again, by the implicit function theorem, we could construct a differentiable curve $\mathbf{x}(t)$ for which $h_i(\mathbf{x}(t)) = 0$ and for which $\mathbf{x}(0) = \mathbf{x}^*$ with $\frac{d}{dt}\mathbf{x}(0) = \mathbf{y}$. Along such a curve $f(\mathbf{x}(t)) = L(\mathbf{x}(t), \boldsymbol{\lambda}^*)$ since $\mathbf{x}(t)$ obeys the constraints. However, using a Taylor series of L, we have

$$\begin{aligned} f(\mathbf{x}(t)) &= L(\mathbf{x}(t), \boldsymbol{\lambda}^*) = L(\mathbf{x}^*, \boldsymbol{\lambda}^*) + (\mathbf{x}(t) - \mathbf{x}^*)^T (\nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*))(\mathbf{x}(t) - \mathbf{x}^*) + o(\|\mathbf{x}(t) - \mathbf{x}^*\|^2) \\ &= f(\mathbf{x}^*) + t^2 \mathbf{y}^T (\nabla^2_{\mathbf{x},\mathbf{x}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*)) \mathbf{y} + o(t^2) \\ &< f(\mathbf{x}^*) \quad \text{for } t \text{ sufficiently close to } 0 \end{aligned}$$

which contradicts the local optimality of $\mathbf{x}^*$.

4.2 Both Inequality and Equality Constraints

Now that we have understood the basis for the necessary conditions for local optimality when there are equality constraints and the associated point is regular, let us augment our focus to include the possibility of inequality constraints and irregularity of the optimum.

We begin by augmenting the problem to include inequality constraints. In particular, we are now interested in optimization problems of the form

$$\min_{\mathbf{x}} \left| \begin{array}{l} f(\mathbf{x}) \\ h_1(\mathbf{x}) = 0,\ h_2(\mathbf{x}) = 0, \ldots,\ h_I(\mathbf{x}) = 0 \\ g_1(\mathbf{x}) \leq 0,\ g_2(\mathbf{x}) \leq 0, \ldots,\ g_J(\mathbf{x}) \leq 0 \end{array} \right. \tag{14}$$

A local minimum of this optimization problem will of course be a point $\mathbf{x}^*$ obeying the constraints, i.e. $h_i(\mathbf{x}^*) = 0, g_j(\mathbf{x}^*) \leq 0,\ \forall i \in \{1,\dots,I\},\ \forall j \in \{1,\dots,J\}$, for which there exists a ball of size $\epsilon > 0$ such that for any $\mathbf{x} \in \mathcal{B}_\epsilon(\mathbf{x}^*)$ such that $h_i(\mathbf{x}) = 0, g_j(\mathbf{x}) \leq 0,\ \forall i \in \{1,\dots,I\},\ \forall j \in \{1,\dots,J\}\ f(\mathbf{x}) \geq f(\mathbf{x}^*)$.

In attempting to reconcile the previous equality constraint case with the new inequality constraints, it is useful to note that for any given $\mathbf{x}^*$, the constraints may be separated in to the set of active constraints, which have indices

$$\mathcal{A}(\mathbf{x}^*) := \{ j \in \{1,\dots,J\} \,|\, g_j(\mathbf{x}^*) = 0 \} \tag{15}$$

and the set of inactive constraints having indices $j \in \{1,\dots,J\} \setminus \mathcal{A}(\mathbf{x}^*)$, for which $g_j(\mathbf{x}^*) < 0$. Because they are inactive and differentiable, there exists a ball around $\mathbf{x}^*$, $\mathcal{B}_{\epsilon'}(\mathbf{x}^*)$, for which any $\mathbf{x} \in \mathcal{B}_{\epsilon'}(\mathbf{x}^*)$ obeys $g_j(\mathbf{x}) < 0$ for all $j \in \{1,\dots,J\} \setminus \mathcal{A}(\mathbf{x}^*)$. Hence, inactivity of these constraints, and the fact that the candidate $\mathbf{x}^*$ obeys them, is sufficient to guarantee that any $\mathbf{x}$ in a small enough neighborhood of $\mathbf{x}^*$ obeys them.

Thus, in considering conditions for local optimality, after verifying that $\mathbf{x}^*$ obeys the constraints, it suffices to consider the constraints that are active. When deriving necessary conditions for local optimality, the active inequality constraints may then be considered as effective equality because if $\mathbf{x}^*$ is a local minimum of (14) must also be a local minimum of

$$\min_{\mathbf{x}} \left| \begin{array}{ll} & f(\mathbf{x}) \\ h_1(\mathbf{x}) = 0,\ h_2(\mathbf{x}) = 0, \dots,\ h_I(\mathbf{x}) = 0 & \\ g_j(\mathbf{x}) = 0,\ j \in \mathcal{A}(\mathbf{x}^*) & \end{array} \right. \tag{16}$$

This is because the latter has a constraint set which is locally a strict subset of the former's constraint set and both contain $\mathbf{x}^*$.

From this we extend the analysis from the previous section to deduce the following necessary conditions for local optimality. In this vein, we define the Lagrangian

$$L(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = f(\mathbf{x}) + \sum_{i=1}^{I} \lambda_i h_i(\mathbf{x}) + \sum_{j=1}^{J} \mu_j g_j(\mathbf{x}) \tag{17}$$

A candidate point for a local minimum $\mathbf{x}^*$ to (14) is said to be *regular* if the gradients of all of the equality constraints and all of the *active* inequality constraints are linearly independent, i.e. if $\{\nabla h_i(\mathbf{x}^*), \nabla g_j(\mathbf{x}^*) \,|\, i \in \{1,\dots,I\}, j \in \mathcal{A}(\mathbf{x}^*)\}$ is a set of linearly independent vectors.

A necessary condition for a *regular* $\mathbf{x}^*$ to be a local minimum of (14) is for there to exist Lagrange multipliers $\lambda_i^*,\ i \in \{1,\dots,I\}$ and $\mu_j^* \geq 0,\ j \in \{1,\dots,J\}$ such that

$$\nabla f(\mathbf{x}^*) + \sum_{i=1}^{I} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{J} \mu_j^* \nabla g_j(\mathbf{x}^*) = \mathbf{0} \quad \text{that is,} \quad \nabla L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*) = \mathbf{0} \tag{18}$$

with $\mu_j^* = 0$ for all $j \notin \mathcal{A}(\mathbf{x}^*)$. Additionally, it is necessary that

$$\mathbf{y}^T (\nabla^2_{\mathbf{xx}} L(\mathbf{x}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*)) \mathbf{y} \geq 0 \tag{19}$$

for all $\mathbf{y}$ such that $(\nabla h_i(\mathbf{x}^*))^T \mathbf{y} = 0$ for all $i \in \{1,\dots,I\}$ and $(\nabla g_j(\mathbf{x}^*))^T \mathbf{y} = 0$ for all $j \in \mathcal{A}(\mathbf{x}^*)$.

These necessary conditions are known as the *Karush Kuhn Tucker conditions* or KKT conditions for local optimality.

Finally, it is of interest to remove the requirement that the local optimum under consideration be regular. In this regard, the *Fritz John conditions* are useful. They change the first order gradient condition to the existence of multipliers $\mu_0^* \geq 0$ and $\lambda_1^*, \dots, \lambda_J^*$ and $\mu_j^* \geq 0$ such that

$$\mu_0^* \nabla f(\mathbf{x}^*) + \sum_{i=1}^{I} \lambda_i^* \nabla h_i(\mathbf{x}^*) + \sum_{j=1}^{J} \mu_j^* \nabla g_j(\mathbf{x}^*) = \mathbf{0} \tag{20}$$

for $\mu_0^*, \lambda_1^*, \dots, \lambda_I^*, \mu_1^*, \dots, \mu_J^*$ not all zero. Additionally, in every neighborhood of $\mathbf{x}^*$ there is an $\mathbf{x}$ such that $\lambda_i^* h_i(\mathbf{x}^*) > 0$ for all i with $\lambda_i^* \neq 0$ and $\mu_j^* g_j(\mathbf{x}^*) > 0$ for all j with $\mu_j^* \neq 0$. (This is actually equivalent to what we have learned so far, since either a point is regular, allowing $\mu_0^* \neq 0$ and allowing us to divide through by it to get the KKT condition, or it is not regular, in which case we can take $\mu_0^* = 0$ and select the other Lagrange multipliers to be a linear combination of the constraint gradients equaling zero, which must exists since the point is not regular.)