Complementary Notes for Nonlinear Optimization (Class 10)

Marcos Raydan*

May 13, 2020

Nonlinear equality constraints: starting at Slide 114

Let us now consider the following constrained optimization problem:

$$\min_{h(x)=0} f(x), \tag{1}$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuously differentiable and its gradient $\nabla f(x)$ is available, $h_i : \mathbb{R}^n \to \mathbb{R}$, $1 \leq i \leq m$, and $h_i \in C^1$ for all i. Note that (1) has m equality constraints but now some of them (at least one) are nonlinear. All the theoretical results that we will discuss can be seen as natural extensions of the results already discussed (NotesClass9) for linear equality constraints.

In the case of linear equality constraints some important facts were automatically satisfied based on some classical linear algebra results. Now, in the presence of nonlinear constraints, the natural extension of the linear case needs some specific additional requirements. For that, we will introduce the concept of a regular point, which allow us to extend the Lagrange theory to obtain first and second order necessary and sufficient optimality conditions. We will also discuss the anomalies that can occur when the possible local minimizers are not regular points. Finally, we will study the numerical algorithms that extend in a natural way Newton-type methods for solving (1), including a suitable option that uses the so-called Schur's complement of the standard 2×2 block matrix associated to the Lagrangian function.

Slides 115, 116, 117: Let us recall that, in the case of linear equality constraints, the feasible set is given by $\Omega = \{x \in \mathbb{R}^n : Ax = b\}$, where $A \in \mathbb{R}^{m \times n}$, $1 \leq m < n$, and $rank(A) = m$. By assuming that $rank(A) = m$ we are assuming that the m rows of A are linearly independent. Note that the rows of A are precisely the gradients of the linear constraints, which are given by $a_{i1}x_1 + a_{i2}x_2 + \cdots + a_{in}x_n = b_i$ for each i.

The natural extension for the case of nonlinear equality constraints is to assume that at a good candidate for a local minimizer, say x^*, the gradient vectors of the constraints are linearly independent. To be precise, $x^* \in \mathbb{R}^n$ is a *regular point* if the vectors $\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$ are linearly independent (LI). Assuming regularity, we are ready to state the first order necessary conditions of problem (1), which extends in a natural way the ones already established for the linear equality constraints problem.

*Centro de Matemática e Aplicações (CMA), FCT, UNL, 2829-516 Caparica, Portugal (`m.raydan@fct.unl.pt`).

Theorem 1. *(First order necessary condition) If x^* is a local minimizer of (1), and x^* is a regular point, then there exist $\lambda_1, \ldots, \lambda_m$ (real numbers) such that*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

In other words, if x^* is a local minimizer of (1) then

$$\nabla f(x^*) \in span\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}.$$

Furthermore, if we extend the Lagrangian function for nonlinear equality constraints:

$$\ell(x, \lambda) = f(x) + \sum_{i=1}^{m} \lambda_i h_i(x) = f(x) + \lambda^T h(x), \tag{2}$$

then we can also say that if x^* is a local minimizer of (1), then

$$\nabla \ell(x^*, \lambda) = \nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

Let us consider a simple example to illustrate the geometrical interpretation of these optimality conditions (see Slides 116 and 117). For that, we consider a problem in which we only have one linear constraint:

$$\min_{x_1^2 + x_2^2 = 2} f(x) = (x_1 + x_2),$$

where $f : \mathbb{R}^2 \to \mathbb{R}$, $\nabla h(x) = 2(x_1, x_2)^T$, and $\nabla f(x) = (1, 1)^T$. By forcing the first order necessary condition: $\nabla f(x) + \lambda \nabla h(x) = 0$, we obtain that $2\lambda x_1 = -1$ and $2\lambda x_2 = -1$, and since $x_1^2 + x_2^2 = 2$, we get that $\lambda^* = 1/2$, and also that $x_1^* = -1$, and $x_2^* = -1$. At x^*, the objective function $f(x^*) = -2$.

Notice that at the solution, $\nabla f(x^*)$ is aligned with the vector $\nabla h(x^*)$ (see Slide 117), i.e., they are parallel vectors (in opposite directions) only differing by a scalar factor. As illustrated in the drawing in Slide 117, this is exactly the interpretation of the first order condition: the point x^* is the one on the feasible set (in this case on the circle given by $x_1^2 + x_2^2 = 2$ in red) for which the corresponding level curve of the objective function (in blue) is tangential to the circle in such a way that the gradient vector at x^* coincides with the gradient of the constraint at the same point x^*, only affected by a scalar factor.

To stress this out, let us consider another vector on the circle (indicated with two arrows emanating from that point, one in blue and one in red). Note that the indicated point is also on the feasible set, and note that at the associated level curve (in blue), the gradient vector points out in a different direction from the gradient vector of the constraint at that point, and hence it cannot be a local solution of the optimization problem. However, it is important to note that there is another point in the feasible set (on the circle) that satisfies the necessary condition: the point $(1, 1)^T$, in which both vectors are also aligned. It also satisfies the first order necessary conditions since $(1, 1)^T$ is the global maximizer.

Slides 118 and 119: For the sake of completeness, we now present the second order optimality conditions of problem (1), whose proofs can be found in any good book in continuous optimization.

Theorem 2. *(Second order necessary condition) If x^* is a local minimizer of (1), and x^* is a regular point, then there exist $\lambda_1, \ldots, \lambda_m$ such that*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0.$$

Moreover, if $f, h \in C^2$,

$$y^T \nabla^2 \ell(x^*, \lambda) y = y^T (\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*)) y \geq 0,$$

for any vector y orthogonal to span$\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$.

Theorem 3. *(Second order sufficient condition) If $h(x^*) = 0$, x^* is a regular point, there exist $\lambda_1, \ldots, \lambda_m$ such that*

$$\nabla f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x^*) = 0,$$

and if

$$y^T \nabla^2 \ell(x^*, \lambda) y = y^T (\nabla^2 f(x^*) + \sum_{i=1}^{m} \lambda_i \nabla^2 h_i(x^*)) y > 0,$$

for any vector $y \neq 0$ orthogonal to span$\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$, then x^* is a strict local minimizer of (1).

Notice that the second order conditions involve the concept of $\nabla^2 \ell(x^*, \lambda)$ being positive (semi) definite on the subspace orthogonal to *span*$\{\nabla h_1(x^*), \ldots, \nabla h_m(x^*)\}$. Hence, the second order conditions can be satisfied even if $\nabla^2 \ell(x^*, \lambda)$ is not positive (semi) definite, as long as it is positive (semi) definite for vectors in that subspace. Notice also that if the inequalities, in Theorems 2 and 3, are written with ≤ 0 and < 0, respectively, then we obtain the second order conditions to find maximizers instead of minimizers.

Let us illustrate the use of the first and second order optimality conditions on the following problem:

$$\text{Optimize} \quad f(x, y) = x^3 + 2y^2 \quad \text{subject to} \quad x^2 + 2y^2 = 4.$$

The associated Lagrangian function is given by:

$$\ell(x, y, \lambda) = x^3 + 2y^2 + \lambda(x^2 + 2y^2 - 4),$$

whose critical points are obtained by solving the nonlinear system ($\nabla \ell(x, y, \lambda) = 0$):

$$\left\{ \begin{array}{rcl} 3x^2 + 2\lambda x & = & 0 \\ 4y + 4\lambda y & = & 0 \\ x^2 + 2y^2 - 4 & = & 0 \end{array} \right. .$$

The second equation can be written as $4y(1 + \lambda) = 0$, and so either $y = 0$ or $\lambda = -1$. If $\lambda = -1$, from the first equation we have that either $x = 0$ or $x = 2/3$. using now the third equation (feasibility) for each possible value of x, we find four critical points: $(0, \sqrt{2})^T$, $(0, -\sqrt{2})^T$, $(2/3, 4/3)^T$, and $(2/3, -4/3)^T$. Now, if $\lambda \neq -1$ then $y = 0$ and we observe from the feasibility that $x \neq 0$, and so from the first equation we obtain $x = -2\lambda/3$. Using now the third equation, we find two additional critical points: $(2,0)^T$ and $(-2,0)^T$. Therefore, there are 6 critical points. Since at any one of them the gradient of the constraint is not the zero vector (i.e., LI), then all of them are regular points, and we can apply the second order optimality conditions.

To determine the global extreme points it is enough to evaluate the objective function in each of the six critical points:

$$f(0,\sqrt{2}) = 4; \quad f(0,-\sqrt{2}) = 4; \quad f(2/3,4/3) = 3.852;$$

$$f(2/3,-4/3) = 3.852; \quad f(2,0) = 8; \quad f(-2,0) = -8.$$

We conclude that $(2,0)^T$ is the global maximizer and $(-2,0)^T$ is the global minimizer.

For the other four critical points we need to use the second order sufficient conditions to classify them. Let us illustrate that process with the point $(0,\sqrt{2})^T$, for which $\lambda = -1$. Evaluating the Hessian of the Lagrangian at that point, we obtain

$$\nabla^2 \ell(x,y,-1) = \begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}.$$

This matrix is negative-semi definite, and we cannot conclude any thing. However, the vectors $w \in \mathbb{R}^2$ orthogonal to the gradient of the constraint at $(0,\sqrt{2})^T$, are characterized by $(w_1,w_2)^T(0,4\sqrt{2}) = 0$ that holds if and only if $4\sqrt{2}w_2 = 0$, i.e., if and only if $w_2 = 0$. Hence, we need to consider for $w_1 \neq 0$:

$$(w_1,0)\begin{bmatrix} -2 & 0 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} w_1 \\ 0 \end{bmatrix} = -2w_1^2 < 0.$$

Hence, that matrix is negative definite on the subspace orthogonal to the gradient of the constraint at the point $(0,\sqrt{2})^T$, and we conclude that it is a local maximizer.

Slides 120 and 121: Let us now discuss some possible Newton-type iterative numerical methods for solving (1). As in the linear equality constraint case, the main motivation comes from considering the Lagrangian function, that in the nonlinear case is given by (2). By differentiating with respect to the vector of variables x and with respect to the vector of Lagrange multipliers λ, we obtain the gradient of $\ell(x,\lambda)$

$$\nabla \ell(x,\lambda) = \begin{bmatrix} \nabla_x \ell(x,\lambda) \\ \nabla_\lambda \ell(x,\lambda) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + J_h(x)\lambda \\ h(x) \end{bmatrix},$$

and the Hessian is given by

$$\nabla^2 \ell(x,\lambda) = \begin{bmatrix} \nabla^2_{xx}\ell(x,\lambda) & J_h(x) \\ J_h(x_k)^T & 0 \end{bmatrix},$$

which is a symmetric and indefinite square block matrix in $\mathbb{R}^{(n+m)\times(n+m)}$ (see Slide 112 to note the similarities with the linear case). In here, the m columns of the $n \times m$ Jacobian matrix $J_h(x)$ are the gradients $\nabla h_i(x)$, $1 \leq i \leq m$, of the nonlinear constraints. Once we have the gradient and the Hessian of the Lagrangian function, we can apply Newton, Quasi-Newton, and low-cost gradient-type methods, to find saddle points of $\ell(x, \lambda)$. Notice that at any critical point x^* of $\ell(x, \lambda)$, the gradient at x^* is zero, i.e., the Lagrange first order necessary condition is satisfied $(\nabla f(x^*) + J_h(x^*)\lambda = 0)$, and also the feasibility is satisfied $(h(x^*) = 0)$. Finally, as in the linear case, note that the block structure of the Hessian is special, including an $m \times m$ zero block in the $(2, 2)$ position.

For Newton's method applied to $\nabla\ell(x, \lambda) = 0$, we need to solve at each iteration the following linear system for $(s_k, \delta_k)^T$ (See slide 113 to compare with the linear equality constraint case):

$$\begin{bmatrix} \nabla^2_{xx}\ell(x_k, \lambda_k) & J_h(x_k) \\ J_h(x_k)^T & 0 \end{bmatrix} \begin{pmatrix} s_k \\ \delta_k \end{pmatrix} = - \begin{pmatrix} \nabla f(x_k) + J_h(x_k)\lambda_k \\ h(x_k) \end{pmatrix}.$$

Once the system is solved, we set $x_{k+1} = x_k + s_k$ and $\lambda_{k+1} = \lambda_k + \delta_k$.

An interesting option for solving the Newton linear system at each iteration k is the so-called Schur's complement approach (explained in Slide 121). This option can also be used in the linear equality constraint case. Since at each iteration, once the blocks at the Hessian matrix (above) are evaluated at x_k and λ_k, they become fixed matrices. Hence, we can present the idea in a simple 2×2 block matrix schematic way, avoiding indices and unnecessary notation.

For that, let us focus on solving for $(s, \delta)^T$ the saddle point linear system (recall the block matrix is symmetric and indefinite):

$$\begin{bmatrix} B & A^T \\ A & 0 \end{bmatrix} \begin{pmatrix} s \\ \delta \end{pmatrix} = \begin{pmatrix} y \\ h \end{pmatrix},$$

where B is square and nonsingular, and A is full row-rank. Apply a 2×2 block Gaussian elimination, i.e., use B as a pivot, multiply the first block row by AB^{-1} and then subtract the second row from the first to obtain the equivalent second row that has a zero block in the $(2, 1)$ position. This elementary Gaussian operation must be done on the left hand side and also on the right hand side of the system, to obtain:

$$\begin{bmatrix} B & A^T \\ 0 & S \end{bmatrix} \begin{pmatrix} s \\ \delta \end{pmatrix} = \begin{pmatrix} y \\ \widehat{h} \end{pmatrix},$$

where $\widehat{h} = h - AB^{-1}y$, and $S = -AB^{-1}A^T$. The matrix $S \in \mathbb{R}^{m \times m}$ is known as the *Schur's complement*. Now, using back substitution, solve for δ the system $S\delta = \widehat{h}$, and then solve for s the system $Bs = y - A^T\delta$.

We can observe that the inverse of the matrix B appears in this approach, which in general is not recommendable. Nevertheless, for problems in which the matrix B (i.e., the Hessian of the Lagrangian with respect to x) is easy to compute (for example, if it is a diagonal or tridiagonal matrix or if it has some specialized structure), then the Schur's complement approach is computationally convenient.

Slides 122 and 123: An issue that needs some additional comments is the regularity at optimal points to be able to guarantee the existence of Lagrange multipliers. Let us recall that in the linear case, this is guaranteed by assuming that the matrix A is full rank. In the nonlinear case some anomalies might occur. First, let us consider (in Slide 122) a simple case in which regularity holds, and hence the existence of Lagrange multipliers is guaranteed. Therefore, we are allowed to use the first and second order conditions. In that example the gradient of the constraint is the vector of all ones, clearly not the zero vector and so it forms a LI set of vectors. The example in Slide 122 is clear and self-explanatory, no more words are needed.

The simple problem described in Slide 123 is more interesting (although it is obviously an academic problem). Notice that the constraints are two circles, one with center in $(-1, 0)$ and radius 1, and the other with center in $(-2, 0)$ and radius 2. Hence, the only feasible point in the intersection of both circles is $(0, 0)$. So, for sure the origin is the minimizer of that problem, or the maximizer if you prefer! In other words, it is very clear that $x^* = (0, 0)^T$ is the solution because it is the only vector in the feasible set. However, as indicated in the slide (easy to check), $\nabla h_1(x^*)$ and $\nabla h_2(x^*)$ are linearly dependent, and $\nabla f(x^*) = (1, 1)^T$ cannot be written as a combination of them (both gradients have a zero in the second position). Consequently, the Lagrange theory is not capable of detecting the trivial solution of this problem. Summing up, regularity is a strong but necessary requirement to apply the discussed necessary and sufficient conditions.

Slide 124: We close this Notes with an important warning in the presence of nonlinear constraints: Watch out with the substitution of variables trying to reduce the number of variables. The potential danger is usually not worth considering the small and most probably insignificant advantage of getting rid of one variable out of our problem. The problem described in Slide 124 is simple and suitable to illustrate this fact.

Let us consider the following nonlinearly constrained optimization problem

$$\min_{(x-1)^3 = y^2} (x^2 + y^2).$$

A clear temptation is to substitute y^2, and consider only one variable and also an unconstrained problem:

$$\min_{x \in \mathbb{R}^n} (x^2 + (x-1)^3).$$

Unfortunately, the original problem has a unique solution $x^* = (1, 0)^T$ (easy to check), but the new (unconstrained) problem has no solutions!

Indeed, for the original constraint problem, the level curves of the objective function are concentric circles with the origin in the center. If we pay attention to the constraint: $(x-1)^3 = y^2$, we note that the term $(x-1)^3$ must be greater than or equal to zero, which implies that $x \geq 1$. That is a hidden (or implicit) additional constraint. So, the feasible set is a curve that starts at the point $(1, 0)^T$ and moves up inside the positive quadrant of $\mathbb{R}^2$. Points in which $x < 1$ or $y < 0$ cannot be valid and so they are not feasible. Hence, the unique global minimizer is $x^* = (1, 0)^T$.

On the other hand, for the unconstrained version, when

$$x \to -\infty, \quad (x^2 + (x-1)^3) \to -\infty,$$

and so there are no minimizers. What really happened when we substitute $y^2 = (x-1)^3$? Answer: the implicit constraint, $x \geq 1$, was gone (forgotten) and the result was catastrophic.