Applied Mathematics 205

Unit IV: Nonlinear Equations and Optimization

Lecturer: Dr. David Knezevic

Unit IV: Nonlinear Equations and Optimization

Chapter IV.3: Conditions for Optimality

Existence of Global Minimum

In order to guarantee existence and uniqueness of a global min. we need to make assumptions about the objective function

e.g. if f is continuous on a closed[1] and bounded set $S \subset \mathbb{R}^n$ then it has global minimum in S

In one dimension, this says f achieves a minimum on the interval $[a, b] \subset \mathbb{R}$

In general f does not achieve a minimum on (a, b), e.g. consider $f(x) = x$

(Though $\inf_{x \in (a,b)} f(x)$, the largest lower bound of f on (a, b), is well-defined)

[1] A set is closed if it contains its own boundary

Existence of Global Minimum

Another helpful concept for existence of global min. is coercivity

A continuous function f on an unbounded set $S \subset \mathbb{R}^n$ is coercive if

$$\lim_{\|x\| \to \infty} f(x) = +\infty$$

That is, $f(x)$ must be large whenever $\|x\|$ is large

Existence of Global Minimum

If f is coercive on a closed, unbounded[2] set S, then f has a global minimum in S

Proof: From the definition of coercivity, for any $M \in \mathbb{R}$, $\exists r > 0$ such that $f(x) \geq M$ for all $x \in S$ where $\|x\| \geq r$

Suppose that $0 \in S$, and set $M = f(0)$

Let $Y \equiv \{x \in S : \|x\| \geq r\}$, so that $f(x) \geq f(0)$ for all $x \in Y$

And we already know that f achieves a minimum (which is at most $f(0)$) on the closed, bounded set $\{x \in S : \|x\| \leq r\}$

Hence f achieves a minimum on S $\square$

[2] e.g. S could be all of $\mathbb{R}^n$, or a "closed strip" in $\mathbb{R}^n$

Existence of Global Minimum

For example:

- $f(x, y) = x^2 + y^2$ is coercive on $\mathbb{R}^2$ (global min. at $(0, 0)$)
- $f(x) = x^3$ is not coercive on $\mathbb{R}$ ($f \to -\infty$ for $x \to -\infty$)
- $f(x) = e^x$ is not coercive on $\mathbb{R}$ ($f \to 0$ for $x \to -\infty$)

Question: What about uniqueness?

Convexity

An important concept for uniqueness is convexity

A set $S \subset \mathbb{R}^n$ is convex if it contains the line segment between any two of its points

That is, S is convex if for any $x, y \in S$, we have

$$\{\theta x + (1-\theta)y : \theta \in [0,1]\} \subset S$$

Convexity

Similarly, we define convexity of a function $f : S \subset \mathbb{R}^n \to \mathbb{R}$

f is convex if its graph along any line segment in S is on or below the chord connecting the function values

i.e. f is convex if for any $x, y \in S$ and any $\theta \in [0, 1]$, we have

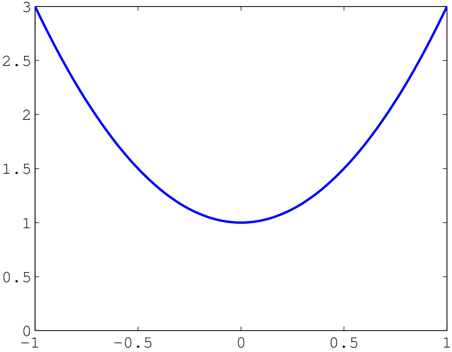
$$f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta)f(y)$$

Also, if

$$f(\theta x + (1 - \theta)y) < \theta f(x) + (1 - \theta)f(y)$$

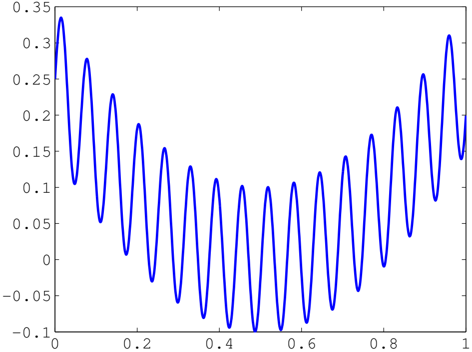
then f is strictly convex

Convexity



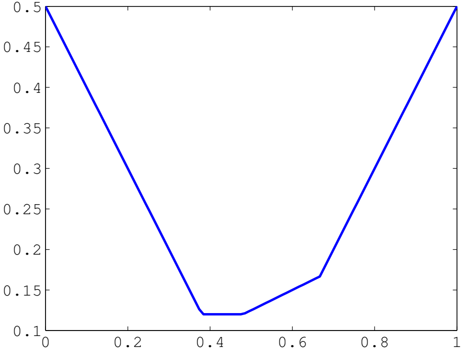
Strictly convex

Convexity



Non-convex

Convexity



Convex (not strictly convex)

Convexity

If f is a convex function on a convex set S, then any local minimum of f must be a global minimum

Proof: Suppose x is a local minimum, i.e. $f(x) \leq f(y)$ for $y \in B(x, \epsilon)$ (where $B(x, \epsilon) \equiv \{y \in S : \|y - x\| \leq \epsilon\}$)

Suppose that x is not a global minimum, i.e. that there exists $w \in S$ such that $f(w) < f(x)$

(Then we will show that this gives a contradiction)

Convexity

Proof (continued...):

For $\theta \in [0,1]$ we have $f(\theta w + (1-\theta)x) \leq \theta f(w) + (1-\theta) f(x)$

Let $\sigma \in (0,1]$ be sufficiently small so that

$$z \equiv \sigma w + (1-\sigma)\, x \in B(x, \epsilon)$$

Then

$$f(z) \leq \sigma f(w) + (1-\sigma)\, f(x) < \sigma f(x) + (1-\sigma)\, f(x) = f(x),$$

i.e. $f(z) < f(x)$, which contradicts that $f(x)$ is a local minimum!

Hence we cannot have $w \in S$ such that $f(w) < f(x)$ $\square$

Convexity

Note that convexity does not guarantee uniqueness of global minimum

e.g. a convex function can clearly have a "horizontal" section (see earlier plot)

If f is a strictly convex function on a convex set S, then a local minimum of f is the unique global minimum

Optimization of convex functions over convex sets is called convex optimization, which is an important subfield of optimization

Optimality Conditions

We have discussed existence and uniqueness of minima, but haven't considered how to find a minimum

The familiar optimization idea from calculus in one dimension is: set derivative to zero, check the sign of the second derivative

This can be generalized to $\mathbb{R}^n$

Optimality Conditions

If $f : \mathbb{R}^n \to \mathbb{R}$ is differentiable, then the gradient vector $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is

$$\nabla f(x) \equiv \begin{bmatrix} \frac{\partial f(x)}{\partial x_1} \\ \frac{\partial f(x)}{\partial x_2} \\ \vdots \\ \frac{\partial f(x)}{\partial x_n} \end{bmatrix}$$

The importance of the gradient is that ∇f points "uphill," i.e. towards points with larger values than $f(x)$

And similarly $-\nabla f$ points "downhill"

Optimality Conditions

This follows from Taylor's theorem for $f : \mathbb{R}^n \to \mathbb{R}$

Recall that

$$f(x + \delta) = f(x) + \nabla f(x)^T \delta + \text{H.O.T.}$$

Let $\delta \equiv -\epsilon \nabla f(x)$ for $\epsilon > 0$ and suppose that $\nabla f(x) \neq 0$, then:

$$f(x - \epsilon \nabla f(x)) \approx f(x) - \epsilon \nabla f(x)^T \nabla f(x) < f(x)$$

Also, we see from Cauchy-Schwarz that $-\nabla f(x)$ is the steepest descent direction

Optimality Conditions

Similarly, we see that a necessary condition for a local minimum at $x^* \in S$ is that $\nabla f(x^*) = 0$

In this case there is no "downhill direction" at x^*

The condition $\nabla f(x^*) = 0$ is called a first-order necessary condition for optimality, since it only involves first derivatives

Optimality Conditions

$x^* \in S$ that satisfies the first-order optimality condition is called a critical point of f

But of course a critical point can be a local min., local max., or saddle point

(Recall that a saddle point is where some directions are "downhill" and others are "uphill", e.g. $(x, y) = (0, 0)$ for $f(x, y) = x^2 - y^2$)

Optimality Conditions

As in the one-dimensional case, we can look to second derivatives to classify critical points

If $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable, then the Hessian is the matrix-valued function $H_f : \mathbb{R}^n \to \mathbb{R}^{n\times n}$

$$H_f(x) \equiv \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n x_1} & \frac{\partial^2 f(x)}{\partial x_n x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix}$$

The Hessian is the Jacobian matrix of the gradient $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$

If the second partial derivatives of f are continuous, then $\partial^2 f/\partial x_i \partial x_j = \partial^2 f/\partial x_j \partial x_i$, and H_f is symmetric

Optimality Conditions

Suppose we have found a critical point x^*, so that $\nabla f(x^*) = 0$

From Taylor's Theorem (see $\text{IV}.2$), for $\delta \in \mathbb{R}^n$, we have

$$\begin{aligned} f(x^* + \delta) &= f(x^*) + \nabla f(x^*)^T \delta + \frac{1}{2}\delta^T H_f(x^* + \eta\delta)\delta \\ &= f(x^*) + \frac{1}{2}\delta^T H_f(x^* + \eta\delta)\delta \end{aligned}$$

for some $\eta \in (0, 1)$

Optimality Conditions

Recall positive definiteness: A is positive definite if $x^T A x > 0$

Suppose $H_f(x^*)$ is positive definite

Then (by continuity) $H_f(x^* + \eta\delta)$ is also positive definite for $\|\delta\|$ sufficiently small, so that: $\delta^T H_f(x^* + \eta\delta)\delta > 0$

Hence, we have $f(x^* + \delta) > f(x)$ for $\|\delta\|$ sufficiently small, i.e. $f(x^*)$ is a local minimum

Hence, in general, positive definiteness of H_f at a critical point x^* is a second-order sufficient condition for a local minimum

Optimality Conditions

A matrix can also be negative definite: $x^T A x < 0$ for all $x \neq 0$

Or indefinite: There exists x, y such that $x^T A x < 0 < y^T A y$

Then we can classify critical points as follows:

- $H_f(x^*)$ positive definite $\implies$ x^* is a local minimum
- $H_f(x^*)$ negative definite $\implies$ x^* is a local maximum
- $H_f(x^*)$ indefinite $\implies$ x^* is a saddle point

Optimality Conditions

Also, positive definiteness of the Hessian is closely related to convexity of f

If $H_f(x)$ is positive definite, then f is convex on some convex neighborhood of x

If $H_f(x)$ is positive definite for all $x \in S$, where S is a convex set, then f is convex on S

Question: How do we test for positive definiteness?

Optimality Conditions

Answer: A is positive (resp. negative) definite if and only if all eigenvalues of A are positive (resp. negative)[3]

Also, a matrix with positive and negative eigenvalues is indefinite

Hence we can compute all the eigenvalues of A and check their signs

[3] This is related to the Rayleigh quotient, see Unit V

Heath Example 6.5

Consider

$$f(x) = 2x_1^3 + 3x_1^2 + 12x_1x_2 + 3x_2^2 - 6x_2 + 6$$

Then

$$\nabla f(x) = \begin{bmatrix} 6x_1^2 + 6x_1 + 12x_2 \\ 12x_1 + 6x_2 - 6 \end{bmatrix}$$

We set $\nabla f(x) = 0$ to find critical points[4] $[1, -1]^T$ and $[2, -3]^T$

[4] In general solving $\nabla f(x) = 0$ requires an iterative method

Heath Example 6.5, continued...

The Hessian is

$$H_f(x) = \begin{bmatrix} 12x_1 + 6 & 12 \\ 12 & 6 \end{bmatrix}$$

and hence

$$\begin{aligned} H_f(1,-1) &= \begin{bmatrix} 18 & 12 \\ 12 & 6 \end{bmatrix}, \text{ which has eigenvalues } 25.4, -1.4 \\ H_f(2,-3) &= \begin{bmatrix} 30 & 12 \\ 12 & 6 \end{bmatrix}, \text{ which has eigenvalues } 35.0, 1.0 \end{aligned}$$

Hence $[2,-3]^T$ is a local min. whereas $[1,-1]^T$ is a saddle point

Optimality Conditions: Equality Constrained Case

So far we have ignored constraints

Let us now consider equality constrained optimization

$$\min_{x \in \mathbb{R}^n} f(x) \quad \text{subject to} \quad g(x) = 0,$$

where $f : \mathbb{R}^n \to \mathbb{R}$ and $g : \mathbb{R}^n \to \mathbb{R}^m$, with $m \leq n$

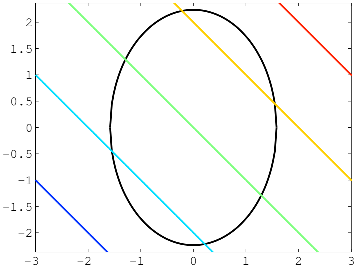
Since g maps to $\mathbb{R}^m$, we have m constraints

This situation is treated with Lagrange mutlipliers

Optimality Conditions: Equality Constrained Case

We illustrate the concept of Lagrange multipliers for $f, g : \mathbb{R}^2 \to \mathbb{R}$

Let $f(x, y) = x + y$ and $g(x, y) = 2x^2 + y^2 - 5$



At any $x \in S$ we must move in direction $(\nabla g(x))_\perp$ to remain in S, hence $(\nabla g(x))_\perp$ is tangent direction[5] (and $\nabla g(x)$ is normal to S)

---

[5] This follows from Taylor's Theorem: $g(x + \delta) \approx g(x) + \nabla g(x)^T \delta$

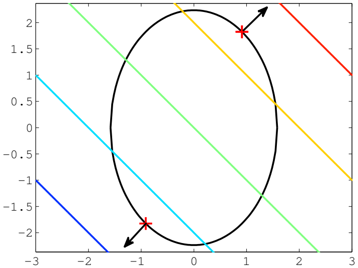
Optimality Conditions: Equality Constrained Case

Also, change in f due to infinitesimal step in direction $(\nabla g(x))_\perp$ is

$$f(x \pm \epsilon(\nabla g(x))_\perp) = f(x) \pm \epsilon \nabla f(x)^T (\nabla g(x))_\perp + \text{H.O.T.}$$

Hence stationary point $x^* \in S$ if $\nabla f(x^*)^T (\nabla g(x^*))_\perp = 0$, or

$$\nabla f(x^*) = \lambda^* \nabla g(x^*), \quad \text{for some } \lambda^* \in \mathbb{R}$$



Optimality Conditions: Equality Constrained Case

This shows that for a stationary point with one constraint, ∇f must be orthogonal to the "tangent direction" of S

Now, consider the case with $m > 1$ equality constraints

Then $g : \mathbb{R}^n \to \mathbb{R}^m$ and we now have a set of constraint gradient vectors, ∇g_i, $i = 1, \ldots, m$

Then we have $S = \{x \in \mathbb{R}^n : g_i(x) = 0, i = 1, \ldots, m\}$

Any "tangent direction" at $x \in S$ must be orthogonal to all gradient vectors $\{\nabla g_i(x), i = 1, \ldots, m\}$ to remain in S

Optimality Conditions: Equality Constrained Case

Let $\mathcal{T}(x) \equiv \{v \in \mathbb{R}^n : \nabla g_i(x)^T v = 0, i = 1, 2, \ldots, m\}$ denote the orthogonal complement of span$\{\nabla g_i(x), i = 1, \ldots, m\}$

Then, for $\delta \in \mathcal{T}(x)$ and $\epsilon \in \mathbb{R}_{>0}$, $\epsilon\delta$ is a step in a "tangent direction" of S at x

Since we have

$$f(x^* + \epsilon\delta) = f(x^*) + \epsilon \nabla f(x^*)^T \delta + \text{H.O.T.}$$

it follows that for a stationary point we need $\nabla f(x^*)^T \delta = 0$ for all $\delta \in \mathcal{T}(x^*)$

Hence at a stationary point $x^* \in S$, $\nabla f(x^*)$ must be in the orthogonal complement of $\mathcal{T}(x^*)$!

Optimality Conditions: Equality Constrained Case

The orthogonal complement of $\mathcal{T}(x^*)$ is span$\{\nabla g_i(x^*), i = 1, \ldots, m\}$, hence:

$$\nabla f(x^*) \in \mathrm{span}\{\nabla g_i(x^*), i = 1, \ldots, m\}$$

This can be written succinctly as a linear system:

$$\nabla f(x^*) = (J_g(x^*))^T \lambda^*$$

for some $\lambda^* \in \mathbb{R}^m$, where $(J_g(x^*))^T \in \mathbb{R}^{n \times m}$

This follows because the columns of $(J_g(x^*))^T$ are the vectors $\{\nabla g_i(x^*), i = 1, \ldots, m\}$

Optimality Conditions: Equality Constrained Case

We can write equality constrained optimization problems more succinctly by introducing the Lagrangian function, $\mathcal{L} : \mathbb{R}^{n+m} \to \mathbb{R}$,

$$\begin{aligned}\mathcal{L}(x,\lambda) &\equiv f(x) + \lambda^T g(x) \\ &= f(x) + \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x)\end{aligned}$$

Then we have,

$$\begin{aligned}\frac{\partial \mathcal{L}(x,\lambda)}{\partial x_i} &= \frac{\partial f(x)}{\partial x_i} + \lambda_1 \frac{\partial g_1(x)}{\partial x_i} + \cdots + \lambda_n \frac{\partial g_n(x)}{\partial x_i}, \quad i = 1,\ldots,n \\ \frac{\partial \mathcal{L}(x,\lambda)}{\partial \lambda_i} &= g_i(x), \quad i = 1,\ldots,m\end{aligned}$$

Optimality Conditions: Equality Constrained Case

Hence

$$\nabla\mathcal{L}(x,\lambda) = \begin{bmatrix} \nabla_x\mathcal{L}(x,\lambda) \\ \nabla_\lambda\mathcal{L}(x,\lambda) \end{bmatrix} = \begin{bmatrix} \nabla f(x) + J_g(x)^T\lambda \\ g(x) \end{bmatrix},$$

so that the first order necessary condition for optimality for the constrained problem can be written as a nonlinear system:[6]

$$\nabla\mathcal{L}(x,\lambda) = \begin{bmatrix} \nabla f(x) + J_g(x)^T\lambda \\ g(x) \end{bmatrix} = 0$$

(As before, stationary points can be classified by considering the Hessian, though we will not consider this here...)

[6] $n+m$ variables, $n+m$ equations

Optimality Conditions: Equality Constrained Case

See Lecture: Constrained optimization of cylinder surface area

Optimality Conditions: Equality Constrained Case

As another example of equality constrained optimization, recall our underdetermined linear least squares problem from I.3

$$\min_{b \in \mathbb{R}^n} f(b) \quad \text{subject to} \quad g(b) = 0,$$

where $f(b) \equiv b^T b$, $g(b) \equiv Ab - y$ and $A \in \mathbb{R}^{m \times n}$ with $m < n$

Optimality Conditions: Equality Constrained Case

Introducing Lagrange multipliers gives

$$\mathcal{L}(b, \lambda) \equiv b^T b + \lambda^T (Ab - y)$$

where $b \in \mathbb{R}^n$ and $\lambda \in \mathbb{R}^m$

Hence $\nabla \mathcal{L}(b, \lambda) = 0$ implies

$$\begin{bmatrix} \nabla f(b) + J_g(b)^T \lambda \\ g(b) \end{bmatrix} = \begin{bmatrix} 2b + A^T \lambda \\ Ab - y \end{bmatrix} = 0 \in \mathbb{R}^{n+m}$$

Optimality Conditions: Equality Constrained Case

Hence, we obtain the $(n+m) \times (n+m)$ square linear system

$$\begin{bmatrix} 2\mathrm{I} & A^T \\ A & 0 \end{bmatrix} \begin{bmatrix} b \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix}$$

We can solve this system analytically for $\begin{bmatrix} b \\ \lambda \end{bmatrix} \in \mathbb{R}^{n+m}$

Optimality Conditions: Equality Constrained Case

We have $b = -\frac{1}{2}A^T\lambda$ from the first "block row"

Subsituting into $Ab = y$ (the second "block row") yields
$\lambda = -2(AA^T)^{-1}y$

And hence

$$b = -\frac{1}{2}A^T\lambda = A^T(AA^T)^{-1}y$$

which was the solution we introduced (but didn't derive) in $\mathrm{I}.3$

Optimality Conditions: Inequality Constrained Case

Similar Lagrange multiplier methods can be developed for the more difficult case of inequality constrained optimization

However, this is outside the scope of AM205...

...though we will use Matlab's functions for inequality constrained optimization