ORF 523	LECTURE 3	Spring 2017, Princeton University
Instructor: A.A. Ahmadi		
Scribe: G. Hall		Tuesday, February 14, 2017

When in doubt on the accuracy of these notes, please cross check with the instructor's notes, on `aaa.princeton.edu/orf523`*. Any typos should be emailed to gh4@princeton.edu.*

Today, we cover the following topics:

- Local versus global minima
- Unconstrained optimization and some of its applications
- Optimality conditions:
  - Descent directions and first order optimality conditions
  - An application: a proof of the arithmetic mean/geometric mean inequality
  - Second order optimality conditions
- Least squares

1 Basic notation and terminology in optimization

1.1 Optimization problems

An optimization problem is a problem of the form

$$\begin{aligned} &\text{min. } f(x) \\ &\text{s.t. } x \in \Omega, \end{aligned} \tag{1}$$

where f is a scalar-valued function called the *objective function*, x is the *decision variable*, and Ω is the *constraint set* (or *feasible set*). The abbreviations min. and s.t. are short for *minimize* and *subject to* respectively. In this class (unless otherwise stated) we always have

$$f : \mathbb{R}^n \to \mathbb{R}, \Omega \subseteq \mathbb{R}^n.$$

Typically, the set Ω is given to us in functional form:

$$\Omega = \{x \in \mathbb{R}^n \mid g_i(x) \geq 0, i = 1, \ldots, m,\ h_j(x) = 0, j = 1, \ldots, k\},$$

for some functions $g_i, h_j : \mathbb{R}^n \to \mathbb{R}$. This is especially the case when we speak of algorithms for solving optimization problems and need explicit access to a description of the set Ω.

1.2 Optimal solution

- An optimal solution x^* (also referred to as the "solution", the "global solution", or the "argmin of f over Ω") is a point in Ω that satisfies:

$$f(x^*) \leq f(x), \forall x \in \Omega.$$

- An optimal solution may not exist or may not be unique.

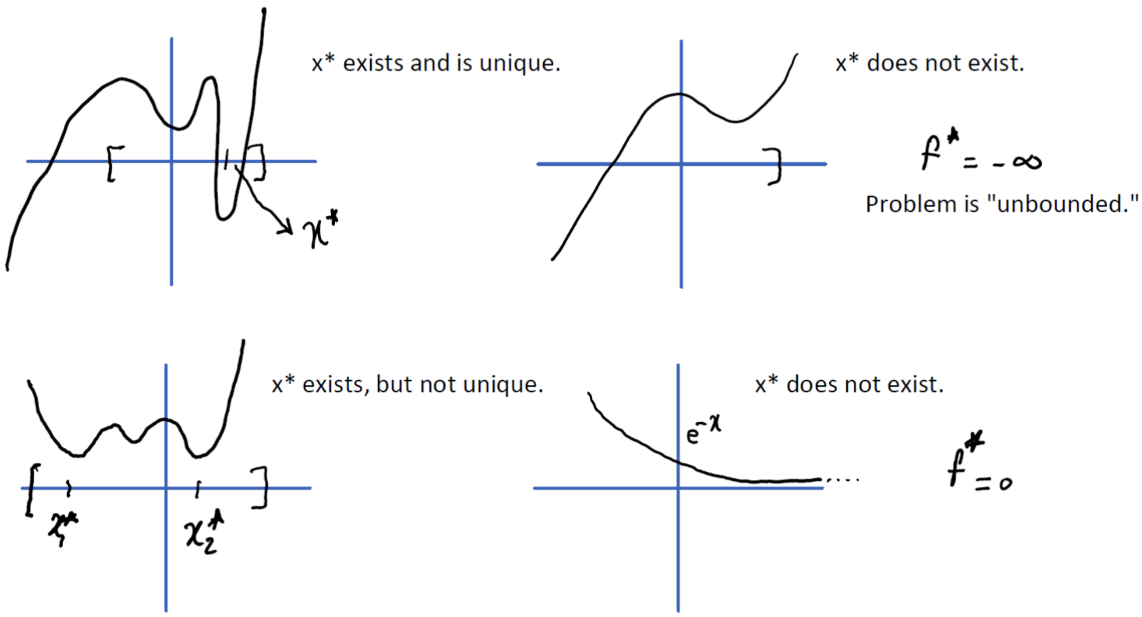


Figure 1: Possibilities for existence and uniqueness of an optimal solution

1.3 Optimal value

- The optimal value f^* of problem (1) is the infimum of f over Ω. If an optimal solution x^* to (1) exists, then the optimal value f^* is simply equal to $f(x^*)$.
- An important case where x^* is guaranteed to exist is when f is continuous and Ω is compact, i.e., closed and bounded. This is known as the Weierstrass theorem. See also Lemma 2 in Section 2.2 for another scenario where the optimal solution is always achieved.
- In the lower right example in Figure 1, the optimal value is zero even though it is not achieved at any x.

- If we want to maximize an objective function instead, it suffices to multiply f by -1 and minimize $-f$. In that case, the optimal solution does not change and the optimal value only changes sign.

1.4 Local and global minima

Consider optimization problem (1). A point $\bar{x}$ is said to be a

- *local minimum*, if $\bar{x} \in \Omega$ and if $\exists \epsilon > 0$ s.t. $f(\bar{x}) \leq f(x),\ \forall x \in B(\bar{x}, \epsilon) \cap \Omega$.
- *strict local minimum* if $\bar{x} \in \Omega$ and if $\exists \epsilon > 0$ s.t. $f(\bar{x}) < f(x),\ \forall x \in B(\bar{x}, \epsilon) \cap \Omega,\ x \neq \bar{x}$.
- *global minimum* if $\bar{x} \in \Omega$ and if $f(\bar{x}) \leq f(x), \forall x \in \Omega$.
- *strict global minimum* if $\bar{x} \in \Omega$ and if $f(\bar{x}) < f(x),\ \forall x \in \Omega,\ x \neq \bar{x}$.

Notation: Here, $B(\bar{x}, \epsilon) := \{x |\ ||x - \bar{x}|| \leq \epsilon\}$. We use the 2-norm in this definition, but any norm would result in the same definition (because of equivalence of norms in finite dimensions).

We can define local/global maxima analogously. Notice that a (strict) global minimum is of course also a (strict) local minimum, but in general finding local minima is a less ambitious goal than finding global minima. Luckily, there are important problems where we can find global minima efficiently.

On the other hand, there are also problems where finding even a local minima is intractable. We will prove the following theorems later in the course:

Theorem 1. *Consider problem (1) with $\Omega = \mathbb{R}^n$. Given a smooth objective function f (even a degree-4 polynomial), and a point $\bar{x}$ in $\mathbb{R}^n$, it is NP-hard to decide if $\bar{x}$ is a local minimum or a strict local minimum of (1).*

Theorem 2. *Consider problem (1) with Ω defined as a set of linear inequalities. Then, given a quadratic function f and a point $\bar{x} \in \mathbb{R}^n$, it is NP-hard to decide if $\bar{x}$ is a local minimum of (1).*

Next, we will see a few optimality conditions that characterize local (and sometimes global) minima. We start with the unconstrained case.

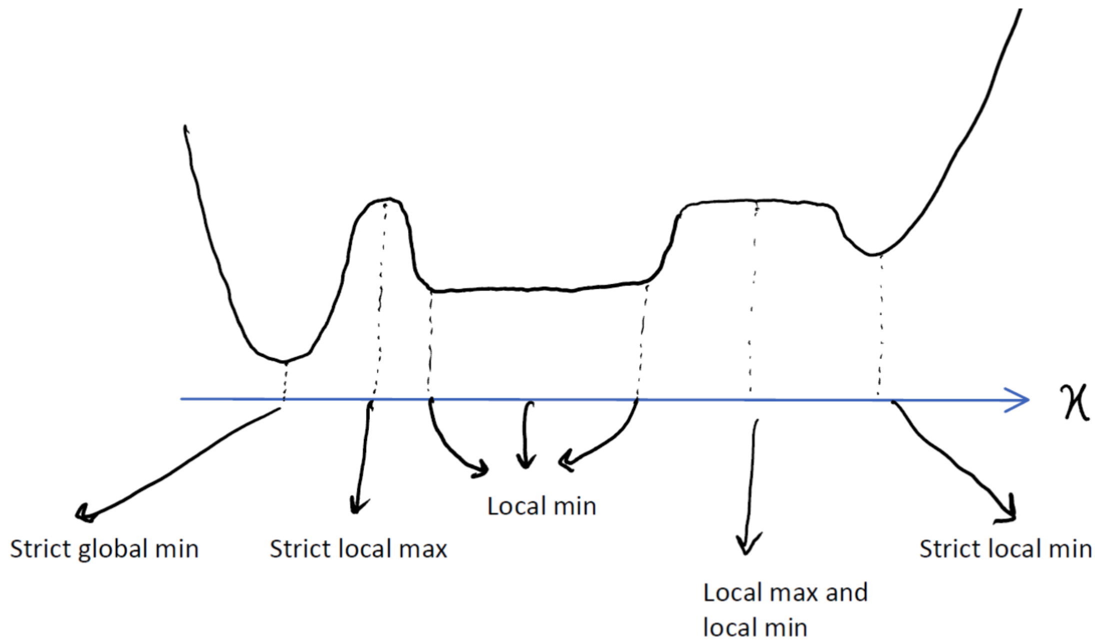


Figure 2: An illustration of local and global minima in the unconstrained case.

2 Unconstrained optimization

Unconstrained optimization corresponds to the case where $\Omega = \mathbb{R}^n$. In other words, the problem under consideration is

$$\min_x f(x).$$

Although this may seem simple, unconstrained problems can be far from trivial. They also appear in many areas of application. Let's see a few.

2.1 Applications of unconstrained optimization

- Example 1: The Fermat-Weber facility location problem. Given locations $z_1, \ldots, z_m$ of households (in $\mathbb{R}^n$), the question is where to place a new grocery store to minimize total travel distance of all customers:

$$\min_{x \in \mathbb{R}^n} \sum_{i=1}^{m} ||x - z_i||$$

- Example 2: Least Squares. There are very few problems that can match least squares in terms of ubiquity of applications. The problem dates back to Gauss: Given $A \in \mathbb{R}^{m \times n}$,

$b \in \mathbb{R}^m$, we are interested in solving the unconstrained optimization problem

$$\min_x ||Ax - b||^2.$$

Typically, $m >> n$. Let us mention a few classic applications of least squares.

- **Data fitting**: We are given a set of points $(x_i, y_i), i = 1, \ldots, N$ on the plane and want to fit a (let's say, degree-3) polynomial $p(x) = c_3x^3 + c_2x^2 + c_1x + c_0$ to this data that minimizes the sum of the squares of the deviations. This, and higher dimensional analogues of it, can be written as a least squares problem (why?).

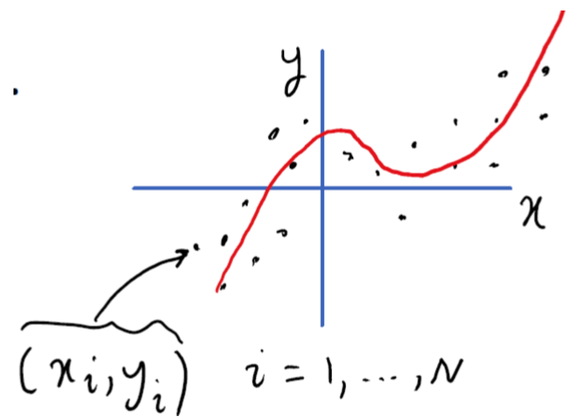


Figure 3: Fitting a curve to a set of data points

- **Overdetermined system of linear equations:** Imagine a very simple linear prediction model for the stock price of a company

$$s(t) = a_1s(t-1) + a_2s(t-2) + a_3s(t-3) + a_4s(t-4),$$

where $s(t)$ is the stock price at day t. We have three months of daily stock price $y(t)$ to train our model. How should we find the best scalars $a_1, \ldots, a_4$ for future prediction? One natural objective is to pick $a_1, \ldots, a_4$ that minimize

$$\sum_{t=1}^{3 \text{ months}} (s(t) - y(t))^2.$$

This is a least squares problem.

- Example 3: Detecting feasibility. Suppose we want to decide if a given set of equalities and inequalities is feasible:

$$S = \{x | \ h_i(x) = 0, i = 1, \ldots, m; \ g_j(x) \geq 0, j = 1, \ldots, k\},$$

where $h_i : \mathbb{R}^n \to \mathbb{R}$, $g_j : \mathbb{R}^n \to \mathbb{R}$. Define

$$f(x, s) = \sum_{i=1}^{m} h_i^2(x) + \sum_{j=1}^{k} (g_j(x) - s_j^2)^2,$$

for some new variables s_j. We have

$$\min_{x,s} f(x, s) = 0 \Leftrightarrow S \text{ is non-empty.}$$

(Why?)

2.2 First order optimality conditions for unconstrained problems

2.2.1 Descent directions

Definition 1. *Consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and a point $x \in \mathbb{R}^n$. A direction $d \in \mathbb{R}^n$ is a descent direction at x if $\exists\ \bar{\alpha} > 0$ s.t.*

$$f(x + \alpha d) < f(x),\ \forall \alpha \in (0, \bar{\alpha}).$$

Lemma 1. *Consider a point $x \in \mathbb{R}^n$ and a continuously differentiable function f. Then, any direction d that satisfies $\nabla^T f(x) d < 0$ is a descent direction. (In particular, $-\nabla f(x)$ is a descent direction if nonzero).*

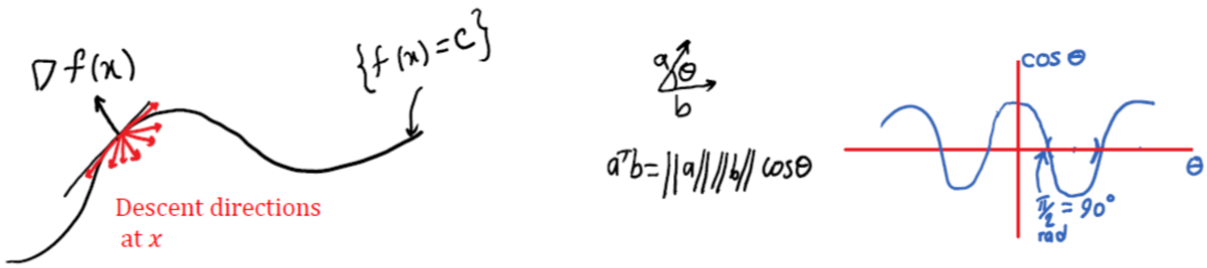


Figure 4: Examples of descent directions

<u>Proof:</u> Let $g : \mathbb{R} \to \mathbb{R}$ be defined as $g(\alpha) = f(x + \alpha d)$ (x and d are fixed here). Then

$$g'(\alpha) = d^T \nabla f(x + \alpha d).$$

We use Taylor expansion to write

$$\begin{aligned} &g(\alpha) = g(0) + g'(0)\alpha + o(\alpha) \\ \Leftrightarrow &f(x+\alpha d) = f(x) + \alpha \nabla^T f(x) d + o(\alpha) \\ \Leftrightarrow &\frac{f(x+\alpha d) - f(x)}{\alpha} = \nabla f^T(x) d + \frac{o(\alpha)}{\alpha} \end{aligned}$$

Since $\lim_{\alpha \downarrow 0} \frac{|o(\alpha)|}{\alpha} = 0$, there exists $\bar{\alpha}$ s.t. $\forall \alpha \in (0, \bar{\alpha})$, we have $\frac{|o(\alpha)|}{\alpha} < \frac{1}{2}|\nabla f^T(x) d|$. Since $\nabla f(x)^T d < 0$ by assumption, we conclude that $\forall \alpha \in (0, \bar{\alpha})$, $f(x+\alpha d) - f(x) < 0$. $\square$

<u>Remark:</u> The converse of Lemma 1 is not true. Consider, e.g., $f(x_1, x_2) = x_1^2 - x_2^2$, $d = (0,1)^T$ and $\bar{x} = (1,0)^T$. For $\alpha \in \mathbb{R}$, we have

$$f(\bar{x} + \alpha d) - f(\bar{x}) = 1^2 - (0 + \alpha^2) - 1^2 + 0^2 = -\alpha^2 < 0,$$

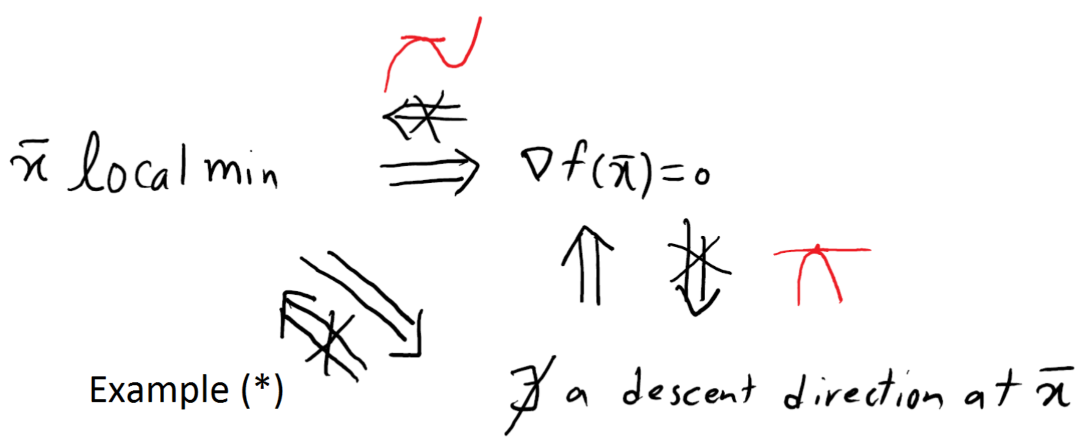
which shows that d is a descent direction for f at $\bar{x}$. But $\nabla^T f(\bar{x}) d = (2,0) \cdot (0,1)^T = 0$.

2.2.2 First order necessary condition for optimality (FONC)

Theorem 3 (Fermat)**.** *If $\bar{x}$ is an unconstrained local minimum of a differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, then $\nabla f(\bar{x}) = 0$.*

<u>Proof:</u> If $\nabla f(\bar{x}) \neq 0$, then $\exists i$ s.t. $\frac{\partial f}{\partial x_i}(\bar{x}) \neq 0$. Then, from Lemma 1, either e_i or $-e_i$ is a descent direction. (Here, e_i is the i^{th} standard basis vector.) Hence, $\bar{x}$ cannot be a local min. $\square$

Let's understand the relationship between the concepts we have seen so far.

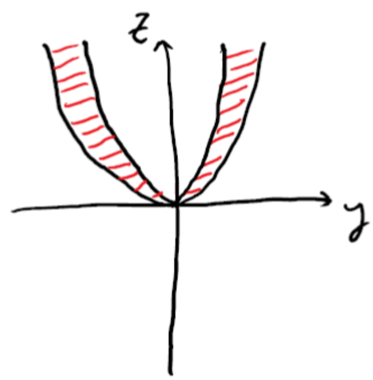


Example (*): Consider the function

$$f(y,z) = (y^2 - z)(2y^2 - z)$$

Claim 1: $(0,0)$ is not a local minimum.
Claim 2: $(0,0)$ is a local minimum along every line that passes through it.

Proof of claim 1: The function $f(y,z) = (y^2 - z)(2y^2 - z)$ is negative whenever $y^2 < z < 2y^2$ and this region gets arbitrarily close to zero; see figure.



Proof of claim 2: For any direction $d = (d_1, d_2)^T$, let's look at $g(\alpha) = f(\alpha d)$:

$$g(\alpha) = (\alpha^2 d_1^2 - \alpha d_2)(2\alpha^2 d_1^2 - \alpha d_2) = 2d_1^4\alpha^4 - 3d_1^2 d_2 \alpha^3 + d_2^2\alpha^2$$
$$g'(\alpha) = 8d_1^4\alpha^3 - 9d_1^2 d_2\alpha^2 + 2d_2^2\alpha$$
$$g''(\alpha) = 24d_1^4\alpha^2 - 18d_1^2 d_2\alpha + 2d_2^2$$
$$g'(0) = 0, g''(0) = 2d_2^2$$

Note that $g'(0) = 0$. Moreover, if $d_2 \neq 0$, $\alpha = 0$ is a (strict) local minimum for g because of the SOSC (see Theorem 5 below). If $d_2 = 0$, then $g(\alpha) = 2d_1^4\alpha^4$ and again $\alpha = 0$ is clearly a (strict) local minimum. □

2.2.3 An application of the first order optimality condition

As an application of the FONC, we give a simple proof of the arithmetic-geometric mean (AMGM) inequality (attributed to Cauchy):

$$(x_1 x_2 \dots x_n)^{1/n} \leq \frac{x_1 + x_2 + \dots + x_n}{n}, \text{ for all } x \geq 0.$$

Our proof follows [1]. We are going to need the following lemma.

Lemma 2. *If a continuous function* $f : \mathbb{R}^n \to \mathbb{R}$ *is* radially unbounded *(i.e.,* $\lim_{||x|| \to \infty} f(x) = \infty$*), then the unconstrained minimum of* f *is achieved.*

Proof: Since $\lim_{||x|| \to \infty} f(x) = \infty$, all sublevel sets of f must be compact (why?). Therefore, $\min_{x \in \mathbb{R}^n} f(x)$ equals

$$\begin{aligned} &\min_x f(x) \\ &\text{s.t. } f(x) \leq \gamma \end{aligned}$$

for any γ for which the latter problem is feasible. Now we can apply Weierstrass and establish the claim. □

Proof of AMGM: The inequality clearly holds if any x_i is zero. So we prove it for $x > 0$. Note that:

$$\begin{aligned} & (x_1 \dots x_n)^{1/n} \leq \frac{\sum_{i=1}^n x_i}{n} \ \forall x > 0 \\ \Leftrightarrow\ & (e^{y_1} \dots e^{y_n})^{1/n} \leq \frac{\sum e^{y_i}}{n} \ \forall y \\ \Leftrightarrow\ & e^{\sum y_i / n} \leq \frac{\sum e^{y_i}}{n} \ \forall y \\ \Leftrightarrow\ & \sum_i e^{y_i} \geq n e^{\sum y_i / n} \ \forall y. \end{aligned} \tag{2}$$

Ideally, we want to show that

$$f(y_1, \dots, y_n) = \sum_i e^{y_i} - n e^{\sum y_i / n} \geq 0, \ \forall y.$$

A possible approach for proving that a function $f : \mathbb{R}^n \to \mathbb{R}$ is nonnegative is to find all points x for which $\nabla f(x) = 0$ and verify that f is nonnegative when evaluated at these points. For this reasoning to be valid though, one needs to be sure that the minimum of f is achieved (see figure below to see why).

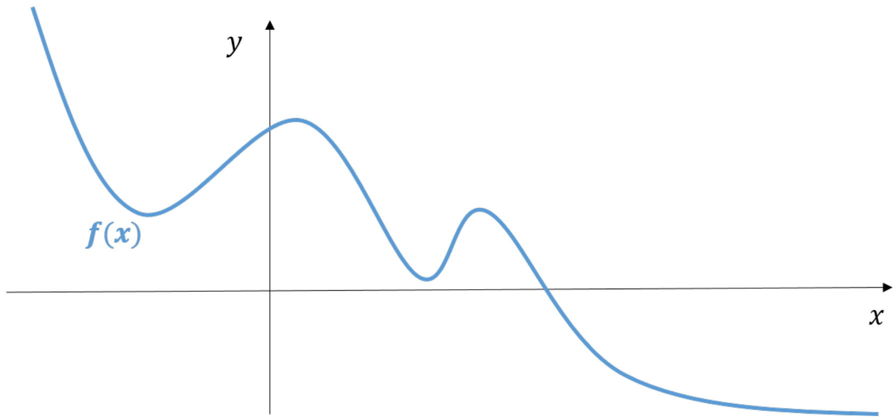


Figure 5: Example of a function f where $f(x) \geq 0$, for all x such that $\nabla f(x) = 0$ without f being nonnegative.

The idea now is to use Lemma (2) to show that the minimum is achieved. But f is not radially unbounded (to see this, take $y_1 = \cdots = y_n$). We will get around this below by working with a function in one less variable that is indeed radially unbounded. Observe that

$$\text{(2) holds} \Leftrightarrow \begin{bmatrix} \min e^{y_1} + \ldots + e^{y_n} \\ \text{s.t. } y_1 + \ldots + y_n = s \end{bmatrix} \geq ne^{s/n} \ \forall s \in \mathbb{R}$$
$$\Leftrightarrow \min e^{y_1} + \ldots + e^{y_{n-1}} + e^{s-(y_1+\ldots+y_{n-1})} \geq ne^{s/n} \ \forall s$$

Define $f_s(y_1, \ldots, y_{n-1}) := e^{y_1} + \ldots + e^{y_{n-1}} + e^{s-y_1-\ldots-y_{n-1}}$. Notice that f_s is radially unbounded (why?). Let's look at the zeros of the gradient of f_s:

$$\frac{\partial f_s}{\partial y_i} = e^{y_i} - e^{s-y_1-\ldots-y_{n-1}} = 0$$
$$\Rightarrow y_i = s - y_1 - \ldots - y_n, \ \forall i$$
$$\Rightarrow y_i^* = \frac{s}{n}, \ i = 1, \ldots, n-1.$$

This is the only solution to $\nabla f_s = 0$. To see this, let's write our equations in matrix form

$$\begin{pmatrix} 2 & 1 & 1 & \ldots & 1 \\ 1 & 2 & 1 & \ldots & 1 \\ \vdots & & \ddots & & \\ \vdots & & & \ddots & \\ 1 & & & & 2 \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ \vdots \\ \vdots \\ y_{n-1} \end{pmatrix} = \begin{pmatrix} s \\ \vdots \\ \vdots \\ \vdots \\ s \end{pmatrix}$$

Denote the matrix on the left by B. Note that $B = 11^T + I \Rightarrow \lambda_{\min}(B) = 1 \Rightarrow det(B) \neq 0$, so the system must have a unique solution.
Now observe that

$$f_s(y^*) = ne^{\frac{s}{n}}.$$

Since $f_s(y^*) = ne^{\frac{s}{n}}$ and f_s is radially unbounded, it follows that

$$f_s(y) \geq ne^{s/n},\ \forall y,$$

and this is true for any s. $\square$

2.3 Second order optimality conditions

2.3.1 Second order necessary and sufficient conditions for local optimality

Theorem 4 (Second Order Necessary Condition for (Local) Optimality (SONC))**.** *If x^* is an unconstrained local minimizer of a twice differentiable function $f : \mathbb{R}^n \to \mathbb{R}$, then in addition to $\nabla f(x^*) = 0$, we must have*

$$\nabla^2 f(x^*) \succeq 0$$

(i.e., the Hessian at x^* is positive semidefinite).

Proof: Consider some $y \in \mathbb{R}^n$. For $\alpha > 0$ the second order Taylor expansion of f around x^* gives

$$f(x^* + \alpha y) = f(x^*) + \alpha y^T \nabla f(x^*) + \frac{\alpha^2}{2} y^T \nabla^2 f(x^*) y + o(\alpha^2).$$

Since $\nabla f(x^*)$ must be zero (as previously proven), we have

$$\frac{f(x^* + \alpha y) - f(x^*)}{\alpha^2} = \frac{1}{2} y^T \nabla^2 f(x^*) y + \frac{o(\alpha^2)}{\alpha^2}.$$

By definition of local optimality of x^*, the left hand side is nonnegative for α sufficiently small. This implies that

$$\lim_{\alpha \downarrow 0} \frac{1}{2} y^T \nabla^2 f(x^*) y + \frac{o(\alpha^2)}{\alpha^2} \geq 0.$$

But

$$\lim_{\alpha \downarrow 0} \frac{o(\alpha^2)}{\alpha^2} = 0 \Rightarrow y^T \nabla^2 f(x^*) y \geq 0.$$

Since y was arbitrary, we must have $\nabla^2 f(x^*) \succeq 0$. $\square$
Remark: The converse of this theorem is not true (why?).

Theorem 5 (Second Order Sufficient Condition for Optimality (SOSC))**.** *Suppose $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable and there exists a point x^* such that $\nabla f(x^*) = 0$, and*

$$\nabla^2 f(x^*) \succ 0$$

(i.e., the Hessian at x^* is positive definite). Then, x^* is a strict local minimum of f.

Proof: Let $\lambda > 0$ be the minimum eigenvalue of $\nabla^2 f(x^*)$. This implies that

$$\begin{aligned}&\nabla^2 f(x^*) - \lambda I \succeq 0\\ &\Rightarrow y^T \nabla^2 f(x^*) y \geq \lambda ||y||^2, \forall y \in \mathbb{R}^n.\end{aligned}$$

Once again, Taylor expansion yields

$$\begin{aligned}f(x^* + y) - f(x^*) &= y^T \nabla f(x^*) + \frac{1}{2} y^T \nabla^2 f(x^*) y + o(||y||^2)\\ &\geq \frac{1}{2}\lambda ||y||^2 + o(||y||^2)\\ &= ||y||^2 \left(\frac{\lambda}{2} + \frac{o(||y||^2)}{||y||^2}\right).\end{aligned}$$

Since $\lim_{||y|| \to 0} \frac{o(||y||^2)}{||y||^2} = 0$, $\exists \delta > 0, \text{s.t. } \frac{o(||y||^2)}{||y||^2} < \frac{\lambda}{2}, \forall y$ with $||y|| \leq \delta$.
Hence,

$$f(x^* + y) > f(x^*), \forall y \text{ with } ||y|| \leq \delta.$$

But this by definition means that x^* is a strict local minimum. □
Remark: The converse of this theorem is not true (why?).

2.3.2 Least squares revisited

Let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$ and suppose that the columns of A are linearly independent. Recall that least squares is the following problem:

$$\text{min. } ||Ax - b||^2.$$

Let $f(x) = ||Ax - b||^2 = x^T A^T A x - 2x^T A^T b + b^T b$. Let's look for candidate solutions among the zeros of the gradient:

$$\nabla f(x) = 2A^T A x - 2A^T b$$

$$\nabla f(x) = 0 \Rightarrow A^T Ax = A^T b \qquad (3)$$
$$\Rightarrow x = (A^T A)^{-1} A^T b$$

Note that the matrix $A^T A$ is indeed invertible because its nullspace is just the origin:

$$A^T Ax = 0 \Rightarrow x^T A^T Ax = 0 \Rightarrow ||Ax||^2 = 0 \Rightarrow Ax = 0 \Rightarrow x = 0,$$

where, for the last implication, we have used the fact that the columns of A are linearly independent. As $\nabla^2 f(x) = 2A^T A \succ 0$ (as $x^T A^T Ax = ||Ax||^2 \geq 0$ and $= 0 \Leftrightarrow x = 0$), then $x = (A^T A)^{-1} A^T b$ is a strict local minimum. Can you argue that x is also the unique *global* minimum? (Hint: Argue that the objective function is radially unbounded and hence the global minimum is achieved.)

2.4 A few remarks to keep in mind

The optimality conditions introduced in this lecture suffer from two problems:

1. It is possible for all three conditions together to be inconclusive about testing local optimality (can you give an example?).

2. They say absolutely nothing about global optimality of solutions.

We will see in the next lecture how to add more structure on f and Ω to get global statements. This will bring us to the fundamental notion of *convexity*.

References

[1] D.P. Bertsekas. *Nonlinear programming, Second Edition.* Athenae Scientific, 2003.