## LINEAR ALGEBRA AND VECTOR ANALYSIS

MATH 22B

# Unit 19: Extrema

## INTRODUCTION

19.1. Learning is an optimization process with the goal to increase knowledge, skills and creative power. This applies both for education as well as for machine learning. In order to track the learning process, we need a function which measures progress. An old fashioned metric is the GPA averaging some grades in an educational system, an other or IQ scores measured by doing tests. An other metric example in a research setting is a social network score like the number of citations or the h-index. For a car driving autonomously it could be the f(x) = 100/(1+N(x)) where N(x) is the number of accidents produced using the parameter configuration x in a fixed period.



Search in a Homing Space: 1. Clueless regions. 2. Large clued regions leading to the target.

Search in a Klondike Space: 1. A large space with few solutions (a wilderness trap). 2. Regions with no clues pointing direction (plateau traps). 3. A barrier isolates the solution (creating a canyon trap). 4. An area of high promise but no solution (an oasis trap).

FIGURE 1. The Klondike picture of David Perkins illustrates the search for solutions in a higher dimensional landscape defined by a height function f. Calculus suggests to follow the gradient  $\nabla f$  as this leads to **local maxima**. To find global maxima (which could be breakthrough ideas), we have to search harder and make a list of all maxima. The process is named after Klondike region in Canada which became infamous during the Gold rush. **19.2.** Once the frame work and the function f is fixed, the question is how to increase f most effectively. This simplistic picture is quite effective both for human intelligence or artificial intelligence. For many functions which have been considered (winning in chess games, computational power, data retention, feature detection, driving cars or flying planes) machines progressed rapidly. There is hardly anybody who seriously doubts that humans eventually will lose the battle for any function f which can be considered. There are still domains where machines have not taken over. Examples are art or writing scientific papers. <sup>1</sup>

**19.3.** Once a machine knows the function f, it can quit comfortably determine from a position x in which direction to change to increase f most rapidly. The direction of **fastest increase** is the direction of the gradient  $\nabla f$  of f. In calculus, we look at situations, where the position consists of a few variables only. Single variable calculus deals with the situation of one variable. We look here at the situation with n variables but will mostly work with 2 variables as this already gives the main idea. The principle is that we have reached an optimum where no change any more can increase the function f. This means mathematically that the derivative df of f is zero. We call such points "critical points".

**19.4.** Let us first look at the rate of change of a function along a direction v. Take a curve r(t) = x + tv where v is a unit vector. By the chain rule, the rate of change at x is given by  $f(r(t)) = \nabla f(r(t)) \cdot r'(0) = \nabla f(x) \cdot v$ . We know for the dot product that this is equal to  $|\nabla f(x)||v| \cos(\alpha) = |\nabla f(x)| \cos(\alpha)$ . This is maximized for  $\cos(\alpha) = 1$  which means that v points into the same direction than  $\nabla f$ . So, The gradient points into the direction of maximal increase. This is important to remember. If you are in a landscape given by the height f(x) you have to go into the direction of  $\nabla f(x)/|\nabla f(x)|$  in order to increase most. Of course, this does not make sense if  $\nabla f(x) = 0$  but that is the situation where you are at a maximum, and where you can not increase f any more.

## Lecture

**19.5.** All functions are assumed here to be in  $C^2$ , meaning that they are two times continuously differentiable. It all starts with an observation going back to Pierre de Fermat:

**Theorem:** If  $x_0$  is a maximum of  $f : \mathbb{R}^m \to \mathbb{R}$ , then  $\nabla f(x_0) = 0$ .

Proof. We prove this by contradiction. Assume  $\nabla f(x_0) \neq 0$ , define the vector  $v = \nabla f(x_0)$  and look at  $g(t) = f(x_0 + tv)$ , which is a function of one variable. By the chain rule, it satisfies  $g'(0) = \nabla f(x_0 + 0v) \cdot v = |\nabla f|^2 > 0$ . This means that  $f(x_0 + tv) > 0$  for small t > 0. The point  $x_0$  can not have been maximal. This is a **contradiction**. QED.

<sup>&</sup>lt;sup>1</sup>There could be resistance: humans might decide not to cite scientific breakthroughs by machines. On the other hand, who would not want to learn a "theory of everything" even if it is discovered by a machine?

**19.6.** A point x with  $\nabla f(x) = 0$  is called a **critical point** of f. By the Taylor formula, we have at a critical point  $x_0$  the quadratic approximation  $Q(x) = f(x_0) + (x - x_0)^T H(x_0)(x - x_0)/2$ , where  $H(x_0)$  is the **Hessian matrix** 

$$H(x_0) = \begin{bmatrix} f_{x_1x_1} & f_{x_1x_2} & \dots & f_{x_1x_m} \\ f_{x_2x_1} & f_{x_2x_2} & \dots & f_{x_2x_m} \\ \dots & \dots & \dots & \dots \\ f_{x_mx_1} & f_{x_mx_2} & \dots & f_{x_mx_m} \end{bmatrix}$$

**19.7.** As in one dimension, having a critical point does not assure that a point is a local maximum or minimum. The second derivative test in single variable calculus assures that if  $f'(x_0) = 0$ ,  $f''(x_0) > 0$ , we have a local minimum and if  $f'(x_0) = 0$ ,  $f''(x_0) < 0$ , we have a local maximum. If  $f''(x_0) = 0$ , we can not say anything without looking at higher derivatives.

**19.8.** A matrix A is called **positive definite** if  $v \cdot Av > 0$  for all vectors  $v \neq 0$ . It is called **negative definite** if  $v \cdot Av < 0$  for all vectors  $v \neq 0$ . A diagonal matrix with positive diagonal entries is positive definite. In the following statements, we assume  $x_0$  is a critical point.

**19.9.** We say  $x_0$  is a **local maximum** of f if there exists r > 0 such that  $f(x) \le f(x_0)$  for all  $|x - x_0| < r$ . We say, it is a **local minimum** of f if  $f(x) \ge f(x_0)$  for all  $|x - x_0| < r$ . How can we check whether a point is a local maximum or minimum?

**Theorem:** Assume  $\nabla f(x_0) = 0$ . If  $H(x_0)$  is positive definite, then  $x_0$  is a local minimum. If  $H(x_0)$  is negative definite, then  $x_0$  is a local maximum.

**19.10.** Proof: as  $\nabla f(x_0) = 0$ , the quadratic approximation at  $x_0$  is  $Q(x) = f(x_0) + H(x_0)v \cdot v/2 > f(x_0)$  for small non-zero  $v = x - x_0$  and Hessian *H*. The analogue statement for the minimum can be deduced by replacing f with -f.

**19.11.** Let us look at the case, where f(x, y) is a function of two variables such that  $f_x(x_0, y_0) = 0$  and  $g_x(x_0, y_0) = 0$ . The Hessian matrix is

$$H(x_0, y_0) = \left[\begin{array}{cc} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{array}\right]$$

In this two dimensional case, we can classify the critical points if the determinant  $D = \det(H) = f_{xx}f_{yy} - f_{xy}^2$  of H is non-zero. The number D is also called the **discriminant** at a critical point.



FIGURE 2.  $f = x^2 + y^2$  gives a minimum,  $f = -x^2 - y^2$  a maximum and  $f = x^2 - y^2$  a saddle. The case  $f = x^2y - yx^2$  is not Morse.

**19.12.** We say  $(x_0, y_0)$  is a **Morse point**, if  $(x_0, y_0)$  is a critical point and the determinant is non-zero. A  $C^2$  function is a **Morse function** if every critical point is Morse. Examples of Morse functions are  $f(x, y) = x^2 + y^2$ ,  $f(x, y) = -x^2 - y^2$  and  $f(x, y) = x^2 - y^2$ . The last case is called a **hyperbolic saddle**. In general, a critical point is a hyperbolic saddle if  $D \neq 0$  and if it is neither a maximum nor a minimum. Here is the **second derivative test** in dimension 2:

**Theorem:** Assume  $f \in C^2$  has a critical point  $(x_0, y_0)$  with  $D \neq 0$ . If D > 0 and  $f_{xx} > 0$  then  $(x_0, y_0)$  is a local minimum. If D > 0 and  $f_{xx} < 0$  then  $(x_0, y_0)$  is a local maximum. If D < 0 then  $(x_0, y_0)$  is a hyperbolic saddle.

**19.13.** Proof. After translation  $(x, y) \to (x - x_0, y - y_0)$  and replacing f with  $f - f(x_0, y_0)$ , we have  $(x_0, y_0) = (0, 0)$  and f(0, 0) = 0. At the critical point, the quadratic approximation is now

$$Q(x,y) = ax^2 + 2bxy + cy^2 .$$

This can be rewritten as  $a(x + \frac{b}{a}y)^2 + (c - \frac{b^2}{a})y^2 = a(A^2 + DB^2)$  with  $A = (x + \frac{b}{a}y), B = b^2/a^2$  and discriminant D. If  $a = f_{xx} > 0$  and D > 0 then  $c - \frac{b^2}{a} > 0$  and the function has positive values for all  $(x, y) \neq (0, 0)$ . The point (0, 0) is then a minimum. If  $a = f_{xx} < 0$  and D > 0, then  $c - \frac{b^2}{a} < 0$  and the function has negative values for all  $(x, y) \neq (0, 0)$ . The point (0, 0) is then a minimum. If  $a = f_{xx} < 0$  and the point (x, y) is a local maximum. If D < 0, then f takes both negative and positive values near (0, 0). QED

**19.14.** One can ask, why  $f_{xx}$  and not  $f_{yy}$  is chosen. It does not matter, because if D > 0, then both  $f_{xx}$  and  $f_{yy}$  need to be non-zero and have the same sign. Instead of  $f_{xx}$ , one could also have pick the more natural **trace** tr(H). It is invariant under coordinate changes similarly as the determinant D. The discriminant D happens also to be the **Gauss curvature** of the surface at the point.

**19.15.** In higher dimensions, the situation is described by the Morse lemma. It tells that near a critical point there is a coordinate change  $\phi$  such that  $g(x) = f(\phi(x))$  is a quadratic function  $f(x) = B(x - x_0) \cdot (x - x_0)$  where B is diagonal with entries +1 or -1. Critical point can then be given a Morse index, the number of entries -1 in B. The Morse lemma is actually a theorem (theorems are more important than lemmata=helper theorems)

**Theorem:** Near a Morse critical point  $x_0$  of a  $C^2$  function f, there is a coordinate change  $\phi : \mathbb{R}^m \to \mathbb{R}^m$  such that  $g(x) = f(\phi(x)) - f(x_0)$  is  $g(x) = -x_1^2 - \cdots - x_k^2 + x_{k+1}^2 + \cdots + x_m^2.$ 

**19.16.** Proof. We use induction with respect to m. (i) Induction foundation. For m = 1, the result tells that for a Morse critical point, the function looks like  $y = x^2$  or  $y = -x^2$ . First show that if f(0) = f'(0) = 0,  $f''(0) \neq 0$ , then  $f(x) = x^2h(x)$  or  $f(x) = -x^2h(x)$  for some positive  $C^2$  function h. Proof. By a linear coordinate change we assume  $x_0 = 0$  and f(0) = 0. There exists then g(x) such that f(x) = xg(x): it is g(x) = f(x)/x for  $x \neq 0$  and is in the limit  $x \to 0$  the value of  $\lim_{x\to 0} (f(x) - f(0))/x = f'(0)$ . By the product rule, f'(x) = g(x) + xg'(x) with g(0) = 0. Because

f'(0) = g(0) = 0 can define  $f(x)/x^2$  for  $x \neq 0$  and take the limit  $x \to 0$ , because by applying Hôpital twice, the limit is f''(0). The coordinate change is now given by a function  $y = \phi(x)$  satisfying  $g(x, y) = y\sqrt{h(y)} = x$ . Implicit differentiation gives  $g_y(0,0) = \sqrt{h(y)} \neq 0$  so that by the implicit function theorem y(x) exists.

(ii) Induction step  $m \to m+1$ : we first note that Taylor for  $C^2$  with remainder term implies that  $f(x_1, \ldots, x_n) = \sum_{i,j} x_i x_j h_{ij}(x_1, \ldots, x_n)$  with some continuous functions  $h_{ij}$ . Furthermore, the function value  $h_{ij}(0) = f_{x_i x_j}(0) = H_{ij}(0)$  are the coordinates of the Hessian. Apply first a rotation so that  $h_{11} \neq 0$ . Now look at  $x_1$  and keep the other coordinates constant. As in (i), find a coordinate change  $\phi$  such that  $f(\phi(x)) =$  $\pm x_1^2 + g(x_2, \ldots, x_m)$ , where g inherits the properties of but is of one dimension less. By induction assumption, there is a second coordinate change such that  $g(\psi(x)) =$  $x_2^2 - \cdots - x_l^2 + x_{l+1}^2 + \cdots + x_m^2$ . Combining  $\phi$  and  $\psi$  produces the Morse normal form.

## EXAMPLES

**19.17.** Q: Classify the critical points of  $f(x, y) = x^3 - 3x - y^3 - 3y$ . A: As  $\nabla f(x, y) = [3x^2 - 3, -3y^2 + 3]^T$ , the critical points are (1, 1), (-1, 1), (1, -1) and (-1, -1). We compute  $H(x, y) = \begin{bmatrix} 2x & 0 \\ 0 & -2y \end{bmatrix}$ . For (1, 1) and (-1, -1) we have D = -4 and so saddle points. For (-1, 1), we have D = 4,  $f_{xx} = -2$ , a local max. For (1, -1) where D = 4,  $f_{xx} = 2$  we have a local min.

## Homework

**Problem 19.1:** a) Classify the critical points of the function

$$f(x,y) = x^2 + y^3 - xy$$

(Maxima, minima or saddle points).

b) Now do the same for

$$f(x, y, z) = x^2 + y^3 - xy + z^2$$

and find the Morse index at each critical point.

Problem 19.2: Find all critical points of the 3D area 51 function  $f(x, y, z) = x^{51} - 51x + y^{51} - 51y + z^{51} - 51z .$ 

Compute the Hessian  $H = d^2 f$  at each critical point and determine the maxima (all eigenvalues are negative) and minima (all eigenvalues are positive) P.S. Area 51 is an old hat. But 3D Area 51 is still highly classified and rumored to be near the dark side of the moon.

**Problem 19.3:** Where on the parametrized surface  $r(u, v) = [u^2, v^3, uv]$  is the temperature T(x, y, z) = 12x + y - 12z minimal. Classify all the critical points of the function f(u, v) = T(r(u, v)). [If you have found the function f(u, v), you can replace u, v again with x, y if you like to work with a function f(x, y).]

**Problem 19.4:** Find all the critical points of the function  $f(x, y, z) = (x-1)^2 - y^2 + xz^2$ . In each of the cases, find the Hessian matrix. Also here compute the eigenvalues. These are numbers  $\lambda$  such that  $Hv = \lambda v$  for some non-zero vector. One can find them by looking for the roots of the characteristic polynomial  $\chi_H(\lambda) = \det(L-\lambda)$ . You can calculate them on a computer. Find in each case the eigenvalues.

**Problem 19.5:** a) Find a function f(x, y) with 3 maxima and 3 saddle points and one minimum.

b) You see below a contour map of a function of two variables. How many critical points are there? Is the function a Morse function?



Oliver Knill, knill@math.harvard.edu, Math 22b, Harvard College, Spring 2022