Calculus Background and Convex Functions

Orestis Plevrakis

In this lecture we will do a review of some topics from Calculus, and we will define convex functions.

1 Calculus

Definition 1. Let $f : A \to \mathbb{R}$, where A is an open subset of $\mathbb{R}^n$. We say that f is C^1 if all the partial derivatives $\frac{\partial f}{\partial x_i}(x)$ exist in A and are continuous functions. We say that f is C^2 if it is C_1, and also all the partial derivatives $\frac{\partial^2 f}{\partial x_i x_j}(x)$ exist in A and are continuous functions.

As you have seen, the gradient of f is

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1}(x), \dots, \frac{\partial f}{\partial x_n}(x) \right)$$

and also the Hessian matrix $\nabla^2 f(x) \in \mathbb{R}^{n \times n}$ has entries

$$\nabla^2 f(x) := \frac{\partial^2 f}{\partial x_i x_j}(x)$$

Clearly, $\nabla f(x)$ and $\nabla^2 f(x)$ are defined wherever the corresponding partial derivatives are defined. From your Calculus class, you know that if f is C^2, then $\nabla^2 f(x) \in S^{n \times n}$ (Clairaut's theorem).

1.1 Simple functions

Here are the simplest multivariate functions:

- Linear: $f(x) = c^\top x$, for some $c \in \mathbb{R}^n$.
- Affine: $f(x) = c^\top x + b$, for some $c \in \mathbb{R}^n, b \in \mathbb{R}$.

Observe that in both cases $\nabla f(x) = c$, $\nabla^2 f(x) = 0$ (zero matrix). Also, affine functions are exactly the polynomials of degree at most one. Here are all the polynomials of degree at most 2:

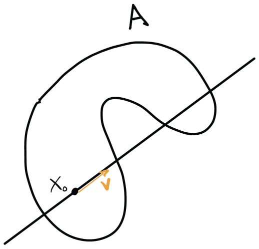
- Quadratic functions: $f(x) = x^\top Ax + c^\top x + b$, for some $A \in S^{n \times n}, c \in \mathbb{R}^n, b \in \mathbb{R}$. To see why these are all the polynomials of degree at most 2, we expand the quadratic form:

  $$x^\top Ax = \sum_i x_i \sum_j A_{ij} x_j = \sum_{i,j} A_{ij} x_i x_j = \sum_i A_{ii} x_i^2 + 2 \sum_{i<j} A_{ij} x_i x_j$$

  Furthermore, $\nabla f(x) = 2Ax + c$ and $\nabla^2 f(x) = 2A$ (check this!).

1.2 Restriction on a line

One of the most important techniques for analyzing multivariate functions is to take their restrictions on lines. These restrictions are functions of one variable, and so we can employ on them all the known machinery. Formally, let $A \subseteq \mathbb{R}^n$ open, $f : A \to \mathbb{R}$. Let x_0 be a point in A and v a vector in $\mathbb{R}^n$. We define $g(t) := f(x_0 + tv)$ where the domain of g is $\text{dom}(g) = \{t \in \mathbb{R} \mid x_0 + tv \in A\}$, which is an open set (why?).



We will study the derivatives of g. Let $t \in \text{dom}(g)$.

- From chain rule, if f is C^1 then g is C^1 and $g'(t) = \nabla f(x_0 + tv) \cdot v$. For $t = 0$, we get

$$\frac{d}{dt} f(x_0 + tv)\bigg|_{t=0} = g'(0) = \nabla f(x_0) \cdot v \tag{1}$$

- If f is C^2, then g is C^2 and

$$\begin{aligned} g''(t) &= \left(\sum_i \frac{\partial f}{\partial x_i}(x_0 + tv) v_i \right)' = \sum_i v_i \nabla \left(\frac{\partial f}{\partial x_i} \right)(x_0 + tv) \cdot v \\ &= \sum_i v_i \sum_j \frac{\partial^2 f}{\partial x_i x_j}(x_0 + tv)\ v_j = \sum_{i,j} \frac{\partial^2 f}{\partial x_i x_j}(x_0 + tv) v_i v_j = v^\top \nabla^2 f(x_0 + tv) v \end{aligned}$$

  For $t = 0$, we get

$$\frac{d^2}{dt^2} f(x_0 + tv)\bigg|_{t=0} = g''(0) = v^\top \nabla^2 f(x_0) v \tag{2}$$

Equations 1 and 2 show us how the gradient and the Hessian matrix encode all the first and second "directional derivatives".

1.3 Fundamental Theorem of Calculus

This is the fundamental theorem of Calculus for functions of one variable: let $f : I \to \mathbb{R}$, where $I \subseteq \mathbb{R}$ is a bounded interval, f is differentiable and has continuous derivative. Let $x, y \in I$. Then, $f(y) - f(x) = \int_x^y f'(u)du = \int_0^1 f'(x_t)dt\ (y - x)$, where $x_t := (1 - t)x + ty$ (in the last step I used change of variables: $t = (u - x)/(y - x)$). For multivariate functions, the theorem generalizes as follows:

Theorem 2. *Let $A \subseteq \mathbb{R}^n$ open, and $f : A \to \mathbb{R}$. Let $x, y \in A$ such that $[x, y] \subseteq A$.*

- *If f is C^1, then $f(y) - f(x) = \int_0^1 \nabla f(x_t)dt \cdot (y - x)$.*

- *If f is C^2, then we also have $\nabla f(y) - \nabla f(x) = \int_0^1 \nabla^2 f(x_t)dt\ (y-x)$*

where $x_t := (1-t)x + ty$,

$$\int_0^1 \nabla f(x_t)dt := \left(\int_0^1 \frac{\partial f}{\partial x_1}(x_t)dt, \dots, \int_0^1 \frac{\partial f}{\partial x_n}(x_t)dt\right)$$

and $\int_0^1 \nabla^2 f(x_t)dt$ is an $n \times n$ matrix whose (i,j) entry is defined as $\left(\int_0^1 \nabla^2 f(x_t)dt\right)_{ij} := \int_0^1 \frac{\partial^2 f}{\partial x_i x_j}(x_t)dt$.

Proof. Suppose f is C_1. Let $g(t) := f(x + t(y-x))$, $t \in [0,1]$. From chain rule, we have that g is differentiable, g' is continuous and $g'(t) = \nabla f(x + t(y-x)) \cdot (y-x)$, and so

$$f(y) - f(x) = g(1) - g(0) = \int_0^1 g'(t)dt = \int_0^1 \nabla f(x_t) \cdot (y-x)dt = \int_0^1 \nabla f(x_t)dt \cdot (y-x)$$

If f is C^2, we can apply the first part of the theorem to $\frac{\partial f}{\partial x_i}$:

$$\frac{\partial f}{\partial x_i}(y) - \frac{\partial f}{\partial x_i}(x) = \int_0^1 \nabla\left(\frac{\partial f}{\partial x_i}\right)(x_t)dt \cdot (y-x) = \int_0^1 \nabla^2 f(x_t)_i\ dt \cdot (y-x)$$

where $\nabla^2 f(x_t)_i$ is the i^{th} row of $\nabla^2 f(x_t)$. This gives the second part. □

1.4 Fermat's theorem and tangent hyperplane

Definition 3. Let $A \subseteq \mathbb{R}^n$, $f : A \to \mathbb{R}$, and $x_0 \in A$.

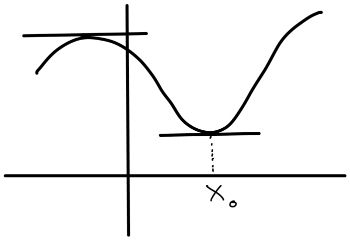
- We say that x_0 is a local minimum if there exists $\epsilon > 0$ such that $f(x) \geq f(x_0)$, for all $x \in B(x_0, \epsilon)$.
- We say that x_0 is a local maximum if there exists $\epsilon > 0$ such that $f(x) \leq f(x_0)$, for all $x \in B(x_0, \epsilon)$.

Theorem 4. *Let $A \subseteq \mathbb{R}^n$ open, $f : A \to \mathbb{R}$, $x_0 \in A$. Suppose f is C^1. If x_0 is local minimum or local maximum, then $\nabla f(x_0) = 0$.*

Proof. Suppose x_0 is local minimum (for local maximum the proof is identical). Let $v \in \mathbb{R}^n$. Since A is open, there exists $\delta > 0$ such that $g(t) = f(x_0 + tv)$ is well-defined for all $t \in (-\delta, \delta)$. Since f is C^1, we have that g is differentiable. Also, $t = 0$ is local minimum for g. Thus, from Fermat's theorem for functions of one variable, we get that $g'(0) = 0$, and so $\nabla f(x_0) \cdot v = 0$. By choosing $v = \nabla f(x_0)$, we are done. □

1.4.1 Geometric interpretation

The geometric interpretation of Fermat's theorem for $n = 1$ is known: if x_0 is a local extremum, then the tangent line of the graph of f at the point $(x_0, f(x_0))$ is parallel to the x axis:



The interpretation for general n is the following: if x_0 is a local extremum, then the "tangent hyperplane" of the "graph of f" at $(x_0, f(x_0))$ is "parallel" to the hyperplane $\{(x,0) \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^{n+1}$. Note that I put " " around the undefined terms. We will define them now, and right after we will verify the correctness of the above statement.

- Parallel hyperplanes: we say that two hyperplanes are parallel if they have parallel normal vectors.
- Graph of a function f: we define it as $G_f := \{(x, f(x)) \mid x \in \text{dom}(f)\} \subseteq \mathbb{R}^{n+1}$.
- Tangent hyperplane: the hyperplane tangent tangent on G_f at $(x_0, f(x_0))$ is

  $$H := \{(x, f(x_0) + \nabla f(x_0) \cdot (x - x_0)) \mid x \in \mathbb{R}^n\}$$

  To see why this is true, check the relevant note at the course website.

Now that we defined everything, let's verify our statement: if x_0 is a local extremum, then Fermat's theorem says that $\nabla f(x_0) = 0$, and so the tangent hyperplane of G_f at $(x_0, f(x_0))$ is $\{(x, f(x_0)) \mid x \in \mathbb{R}^n\}$, which is parallel to $\{(x,0) \mid x \in \mathbb{R}^n\}$ since they are both normal to $(0,\ldots,0,1)$.

2 Convex functions

Definition 5. Let $f : K \to \mathbb{R}$, where K is a convex subset of $\mathbb{R}^n$. We say that f is convex if for any $x, y \in K$ we have

$$\forall \theta \in [0,1], \quad f((1-\theta)x + \theta y) \leq (1-\theta)f(x) + \theta f(y) \tag{3}$$

Geometric interpretation. Let's fix two points $x, y \in K$. We will now show that 3 means that the chord connecting $(x, f(x))$ and $(y, f(y))$ is above G_f.

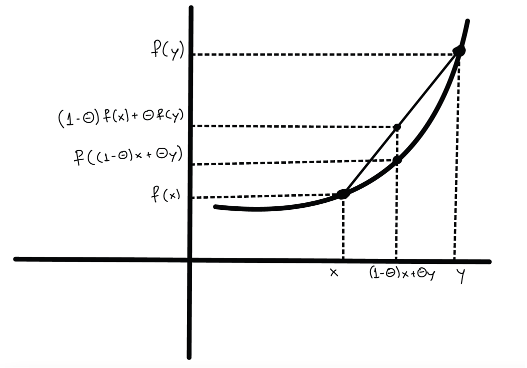


Figure 1: Geometric interpretation for $n = 1$.

In $\mathbb{R}^{n+1}$, the chord is

$$\{(1-\theta)(x, f(x)) + \theta(y, f(y)) \mid \theta \in [0,1]\} = \{((1-\theta)x + \theta y, (1-\theta)f(x) + \theta f(y)) \ \mid \theta \in [0,1]\}$$

and the corresponding part of the graph of f is $\{((1-\theta)x + \theta y, f((1-\theta)x + \theta y)) \ \mid \theta \in [0,1]\}$, and so 3 exactly says that the chord is above the graph.