Lecture 2

In this lecture we will study the unconstrained problem

$$\text{minimize} \quad f(\boldsymbol{x}), \tag{2.1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$. Optimality conditions aim to identify properties that potential minimizers need to satisfy in relation to $f(\boldsymbol{x})$. We will review the well known local optimality conditions for differentiable functions from calculus. We then introduce convex functions and discuss some of their properties.

2.1 Unconstrained optimization

Solutions to (2.1) come in different flavours, as in the following definition.

Definition 2.1. A point $\boldsymbol{x}^* \in \mathbb{R}^n$ is a

- *global minimizer* of (2.1) if for all $\boldsymbol{x} \in \mathbb{R}^n$, $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$;
- a *local minimizer*, if there is an open neighbourhood U of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$;
- a *strict local minimizer*, if there is an open neighbourhood U of $\boldsymbol{x}^*$ such that $f(\boldsymbol{x}^*) < f(\boldsymbol{x})$ for all $\boldsymbol{x} \in U$;
- an *isolated minimizer* if there is an open neighbourhood U of $\boldsymbol{x}^*$ such that $\boldsymbol{x}^*$ is the only local minimizer in U.

Without any further assumptions on f, finding a minimizer is a hopeless task: we simply can't examine the function at *all* points in $\mathbb{R}^n$. The situation becomes more tractable if we assume some *smoothness* conditions. Recall that $C^k(U)$ denotes the set of functions that are k times continuously differentiable on some set U. The following *first-order* necessary condition for optimality is well known. We write $\nabla f(\boldsymbol{x})$ for the gradient of f at $\boldsymbol{x}$, i.e., the vector

$$\nabla f(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_1}(\boldsymbol{x}), \ldots, \frac{\partial f}{\partial x_n}(\boldsymbol{x}) \right)^\top$$

Theorem 2.2. *Let $\boldsymbol{x}^*$ be a local minimizer of f and assume that $f \in C^1(U)$ for a neighbourhood of U of $\boldsymbol{x}^*$. Then $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$.*

There are simple examples that show that this is not a sufficient condition: maxima and saddle points will also have a vanishing gradient. If we have access to *second-order information*, in form of the second derivative, or Hessian, of f, then we can say more. Recall that the Hessian of f at $\boldsymbol{x}$, $\nabla^2 f(\boldsymbol{x})$, is the $d \times d$ symmetric matrix given by the second derivatives,

$$\nabla^2 f(\boldsymbol{x}) = \left(\frac{\partial^2 f}{\partial x_i \partial x_j} \right)_{1 \le i,j \le n}.$$

In the one-variable case we have learned that if x^* is a local minimizer of $f \in C^2([a,b])$, then $f'(x^*) = 0$ *and* $f''(x^*) \ge 0$. Moreover, the conditions $f'(x^*) = 0$ *and* $f''(x^*) > 0$ guarantee that we have a local minimizer. These conditions generalise to higher dimension, but first we need to know what $f''(x) > 0$ when we have more than one variable.

Recall also that a matrix $\boldsymbol{A}$ is **positive semidefinite**, written $\boldsymbol{A} \succeq \mathbf{0}$, if for every $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} \ge 0$, and positive definite, written $\boldsymbol{A} \succ \mathbf{0}$, if $\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0$. The property that the Hessian matrix is positive semidefinite is a multivariate generalization of the property that the second derivative is nonnegative. The known conditions for a minimizer involving the second derivative generalize accordingly.

Theorem 2.3. *Let $f \in C^2(U)$ for some open set U and $\boldsymbol{x}^* \in U$. If $\boldsymbol{x}^*$ is a local minimizer, then $\nabla f(\boldsymbol{x}^*) = 0$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive semidefinite. Conversely, if $\nabla f(\boldsymbol{x}^*) = \mathbf{0}$ and $\nabla^2 f(\boldsymbol{x}^*)$ is positive definite, then $\boldsymbol{x}^*$ is a strict local minimizer.*

Unfortunately, the above criteria are not able to identify global minimizers, as differentiability is a local property.

2.2 Convex functions

We now come to the central notion of this course.

Definition 2.4. A set $C \subseteq \mathbb{R}^n$ is **convex** if for all $\boldsymbol{x}, \boldsymbol{y} \in C$ and $\lambda \in [0,1]$, the line $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in C$. A **convex body** is a convex set that is closed and bounded.

Definition 2.5. Let $S \subseteq \mathbb{R}^n$. A function $f\colon S \to \mathbb{R}$ is called *convex* if S is convex and for all $\boldsymbol{x}, \boldsymbol{y} \in S$ and $\lambda \in [0,1]$,

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \le \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

The function f is called *strictly convex* if

$$f(\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y}) < \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$

A function f is called *concave*, if $-f$ is convex.

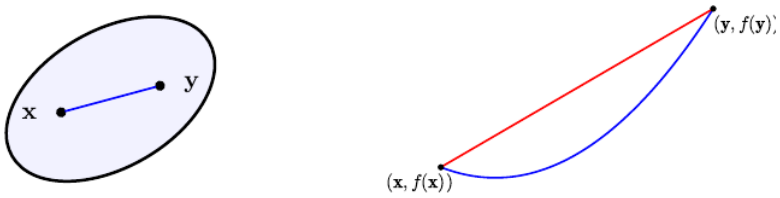

Figure 2.1: A convex set and a convex function

Figure 2.1 illustrates how a convex function of one variable looks like. The graph of the function lies below any line connecting two points on it.

Convex function have pleasant properties, while at the same time covering many of the functions that arise in applications. Perhaps the most important property is that local minima are global minima.

Theorem 2.6. *Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be a convex function. Then any local minimizer of f is a global minimizer.*

Proof. Let $\boldsymbol{x}^*$ be a local minimizer and assume that it is not a global minimizer. Then there exists a vector $\boldsymbol{y} \in \mathbb{R}^d$ such that $f(\boldsymbol{y}) < f(\boldsymbol{x}^*)$. Since f is convex, for any $\lambda \in [0,1]$ and $\boldsymbol{x} = \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{x}^*$ we have

$$f(\boldsymbol{x}) \leq \lambda f(\boldsymbol{y}) + (1-\lambda) f(\boldsymbol{x}^*) < \lambda f(\boldsymbol{x}^*) + (1-\lambda) f(\boldsymbol{x}^*) = f(\boldsymbol{x}^*).$$

This holds for all $\boldsymbol{x}$ on the line segment connecting $\boldsymbol{y}$ and $\boldsymbol{x}^*$. Since every open neighbourhood U of $\boldsymbol{x}^*$ contains a bit of this line segment, this means that every open neighbourhood U of $\boldsymbol{x}^*$ contains an $\boldsymbol{x} \neq \boldsymbol{x}^*$ such that $f(\boldsymbol{x}) \leq f(\boldsymbol{x}^*)$, in contradiction to the assumption that $\boldsymbol{x}^*$ is a local minimizer. It follows that $\boldsymbol{x}^*$ has to be a global minimizer. □

Remark 2.7. Note that in the above theorem we made no assumptions about the differentiability of the function f! In fact, while a convex function is always *continuous*, it need not be differentiable. The function $f(x) = |x|$ is a typical example: it is convex, but not differentiable at $x = 0$.

Example 2.8. Affine functions $f(\boldsymbol{x}) = \langle \boldsymbol{x}, \boldsymbol{a} \rangle + b$ and the exponential function e^x are examples of convex functions.

Example 2.9. In optimization we will often work with functions of matrices, where an $m \times n$ matrix is considered as a vector in $\mathbb{R}^{m \times n} \cong \mathbb{R}^{mn}$. If the matrix is symmetric, that is, if $\boldsymbol{A}^\top = \boldsymbol{A}$, then we only care about the upper diagonal entries, and we consider the space $\mathcal{S}^n$ of symmetric matrices as a vector space of dimension $n(n+1)/2$ (the number of entries on and above the main diagonal). Important functions on symmetric matrices that are convex are the operator norm $\|\boldsymbol{A}\|_2$, defined as

$$\|\boldsymbol{A}\|_2 := \max_{\boldsymbol{x}\colon \|\boldsymbol{x}\| \leq 1} \frac{\|\boldsymbol{A}\boldsymbol{x}\|_2}{\|\boldsymbol{x}\|_2},$$

or the function $\log\det(\boldsymbol{X})$, defined on the set of *positive semidefinite* symmetric matrices $\mathcal{S}_+^n$.

There are useful ways of characterising convexity using differentiability.

Theorem 2.10. *1. Let $f \in C^1(\mathbb{R}^n)$. Then f is convex if and only if for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$,*

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}).$$

2. Let $f \in C^2(\mathbb{R}^n)$. Then f is convex if and only if $\nabla^2 f(\boldsymbol{x})$ is positive semidefinite. If $\nabla^2 f(\boldsymbol{x})$ is positive definite, then f is strictly convex.

Example 2.11. Consider a quadratic function of the form

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^\top \boldsymbol{x} + c,$$

where $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is symmetric. Writing out the product, we get

$$\begin{aligned}\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} &= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \\ &= \begin{pmatrix} x_1 & \cdots & x_n \end{pmatrix} \begin{pmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{pmatrix} = \sum_{i=1}^n \sum_{j=1}^n a_{ij}x_i x_j.\end{aligned}$$

Because $\boldsymbol{A}$ is symmetric, we have $a_{ij} = a_{ji}$, and the above product simplifies to

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^n a_{ii}x_i^2 + 2\sum_{1\leq i<j\leq n} a_{ij}x_i x_j.$$

This is a quadratic function, because it involves products of the x_i. The gradient and the Hessian of $f(\boldsymbol{x})$ are found by computing the partial derivatives of f:

$$\frac{\partial f}{\partial x_i} = \sum_{j=1}^n a_{ij}x_j + b_i, \quad \frac{\partial^2 f}{\partial x_i \partial x_j} = a_{ij}.$$

In summary, we have

$$\nabla f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \quad \nabla^2 f(\boldsymbol{x}) = \boldsymbol{A}.$$

Using the previous theorem, we see that f is convex **if and only if** $\boldsymbol{A}$ is positive semidefinite. A typical example for such a function is

$$f(\boldsymbol{x}) = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}\|^2 = (\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b})^T(\boldsymbol{A}\boldsymbol{x} - \boldsymbol{b}) = \boldsymbol{x}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} - 2\boldsymbol{b}^T\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}^T\boldsymbol{b}.$$

The matrix $\boldsymbol{A}^T\boldsymbol{A}$ is always symmetric and positive semidefinite (why?) so that the function f is convex.

A convenient way to visualise a function $f\colon \mathbb{R}^2 \to \mathbb{R}$ is through **contour plots**. A **level set** of the function f is a set of the form

$$\{\boldsymbol{x} : f(\boldsymbol{x}) = c\},$$

where c is the **level**. Each such level set is a curve in $\mathbb{R}^2$, and a contour plot is a plot of a collection of such curves for various c. If one colours the areas between adjacent curves, one gets a plot as in the following figure. A *convex function* is has the property that there is only one *sink* in the contour plot.

In [1]:

```
# import numpy as np
import numpy.linalg as la
import matplotlib.pyplot as plt
% matplotlib inline

# Create random data: we use the randn function
X = np.random.randn(3,2)
y = np.random.randn(3)

# Solve least squares problem minimize \|X\beta-y\|^2
# the index 0 says that we get the first component of the solution
# (the function lstsq give more output than just the beta vector)
beta = la.lstsq(X,y)[0]

# Create function and plot the contours
def f(a,b):
    return sum((a*X[:,0]+b*X[:,1]-y)**2)

# Find the "right" boundaries around the minimum
xx = np.linspace(beta[0]-8,beta[0]+8,100)
yy = np.linspace(beta[1]-8,beta[1]+8,100)
XX, YY = np.meshgrid(xx,yy)

Z = np.zeros(XX.shape)
for i in range(Z.shape[0]):
    for j in range(Z.shape[1]):
        Z[i,j] = f(XX[i,j],YY[i,j])

cmap = plt.cm.get_cmap("coolwarm")
plt.contourf(XX,YY,Z, cmap = cmap)
plt.show()
```