

---

# Lecture 3

---

Most modern optimization methods are iterative: they generate a sequence of points  $\mathbf{x}_0, \mathbf{x}_1, \dots$  in  $\mathbb{R}^n$  in the hope that this sequence will converge to a local or global minimizer  $\mathbf{x}^*$  of a function  $f(\mathbf{x})$ . A typical rule for generating such a sequence would be to start with a vector  $\mathbf{x}_0$ , chosen by an educated guess, and then for  $k \geq 0$ , move from step  $k$  to  $k + 1$  by

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \alpha_k \mathbf{p}_k,$$

in a way that ensures that  $f(\mathbf{x}_{k+1}) \leq f(\mathbf{x}_k)$ . The parameter  $\alpha_k$  is called the **step length**, while  $\mathbf{p}_k$  is the **search direction**. In this lecture we discuss one such method, the method of gradient descent, or steepest descent, and discuss how to select the right step length.

## 3.1 Gradient descent

In the method of gradient descent, the search direction is chosen as

$$\mathbf{p}_k = -\nabla f(\mathbf{x}_k). \quad (3.1)$$

To see why this makes sense, let  $\mathbf{p}$  be a direction and consider the Taylor expansion

$$f(\mathbf{x}_k + \alpha \mathbf{p}) = f(\mathbf{x}_k) + \alpha \langle \mathbf{p}, \nabla f(\mathbf{x}_k) \rangle + O(\alpha^2).$$

Considering this as a function of  $\alpha$ , the rate of change in direction  $\mathbf{p}$  at  $\mathbf{x}_k$  is the derivative of this function at  $\alpha = 0$ ,

$$\left. \frac{df(\mathbf{x}_k + \alpha \mathbf{p})}{d\alpha} \right|_{\alpha=0} = \langle \mathbf{p}, \nabla f(\mathbf{x}_k) \rangle,$$

also known as the **directional derivative** of  $f$  at  $\mathbf{x}_k$  in the direction  $\mathbf{p}$ . This formula indicates that the rate of change is *negative*, and we have a **descent direction**, if  $\langle \mathbf{p}, \nabla f(\mathbf{x}_k) \rangle < 0$ .

The Cauchy-Schwarz inequality (see Preliminaries, Page 9) gives the bounds

$$-\|\mathbf{p}\|_2 \|\nabla f(\mathbf{x}_k)\|_2 \leq \langle \mathbf{p}, \nabla f(\mathbf{x}_k) \rangle \leq \|\mathbf{p}\|_2 \|\nabla f(\mathbf{x}_k)\|_2.$$

We see that the rate of change is the smallest when the first inequality is an equality, which happens if

$$\mathbf{p} = -\alpha \nabla f(\mathbf{x}_k)$$

for some  $\alpha > 0$ .

For a visual interpretation of what it means to be a descent direction, note that the **angle**  $\theta$  between a vector  $\mathbf{p}$  and the gradient  $\nabla f(\mathbf{x})$  at a point  $\mathbf{x}$  is given by (see Preliminaries, Page 9)

$$\langle \mathbf{x}, \nabla f(\mathbf{x}) \rangle = \|\mathbf{p}\|_2 \|\nabla f(\mathbf{x})\|_2 \cos(\theta).$$

This is negative if the vector  $\mathbf{p}$  forms an angle greater than  $\pi/2$  with the gradient. Recall that the gradient points in the direction of steepest ascent, and is orthogonal to the *level sets*. If you are standing on the slope of a mountain, walking along the level set lines will not change your elevation, the gradient points to the steepest upward direction, and the negative gradient to the steepest descent.

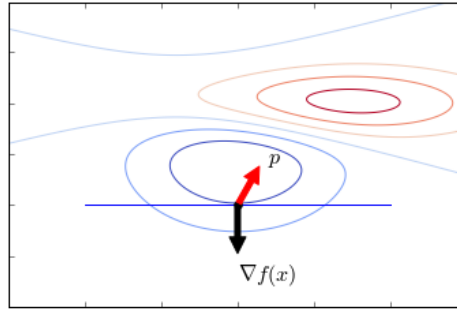


Figure 3.1: A descent direction

Any multiple  $\alpha \nabla f(\mathbf{x}_k)$  points in the direction of steepest descent, but we have to choose a sensible parameter  $\alpha$  to ensure that we make sufficient progress, but at the same time don't overshoot. Ideally, we would choose the value  $\alpha_k$  that minimizes  $f(\mathbf{x}_k - \alpha_k \nabla f(\mathbf{x}_k))$ . While finding such a minimizer is in general not easy (see Section 3.2 for alternatives), for quadratic functions it can be given in closed form.

### Linear least squares

Consider a function of the form

$$f(\mathbf{x}) = \frac{1}{2} \|\mathbf{Ax} - \mathbf{b}\|_2^2.$$

In Problem Sheet 1 you will show that the Hessian is symmetric and positive semidefinite, with the gradient given by

$$\nabla f(\mathbf{x}) = \mathbf{A}^\top (\mathbf{Ax} - \mathbf{b}).$$

The method of gradient descent proceeds as

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_k \mathbf{A}^\top (\mathbf{A}\mathbf{x}_k - \mathbf{b}).$$

To find the best  $\alpha_k$ , we compute the minimum of the function

$$\alpha \mapsto f(\mathbf{x}_k - \alpha \mathbf{A}^\top (\mathbf{A}\mathbf{x}_k - \mathbf{b})). \quad (3.2)$$

If we set  $\mathbf{r}_k := \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}_k) = -\nabla f(\mathbf{x}_k)$  and compute the minimum of (3.2) by differentiating, we get the step length

$$\alpha_k = \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A}^\top \mathbf{A} \mathbf{r}_k} = \frac{\|\mathbf{r}_k\|_2^2}{\|\mathbf{A}\mathbf{r}_k\|_2^2}.$$

(Verify this!) Note also that when we have  $\mathbf{r}_k$  and  $\alpha_k$ , we can compute the next  $\mathbf{r}_k$  as

$$\begin{aligned} \mathbf{r}_{k+1} &= \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}_{k+1}) \\ &= \mathbf{A}^\top (\mathbf{b} - \mathbf{A}(\mathbf{x}_k + \alpha_k \mathbf{r}_k)) \\ &= \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}_k - \alpha_k \mathbf{A}^\top \mathbf{A} \mathbf{r}_k) = \mathbf{r}_k - \alpha_k \mathbf{A}^\top \mathbf{A} \mathbf{r}_k. \end{aligned}$$

The gradient descent algorithm for the linear least squares problem proceeds by first computing  $\mathbf{r}_0 = \mathbf{A}^\top (\mathbf{b} - \mathbf{A}\mathbf{x}_0)$ , and then at each step

$$\begin{aligned} \alpha_k &= \frac{\mathbf{r}_k^\top \mathbf{r}_k}{\mathbf{r}_k^\top \mathbf{A}^\top \mathbf{A} \mathbf{r}_k} \\ \mathbf{x}_{k+1} &= \mathbf{x}_k + \alpha_k \mathbf{r}_k \\ \mathbf{r}_{k+1} &= \mathbf{r}_k - \alpha_k \mathbf{A}^\top \mathbf{A} \mathbf{r}_k. \end{aligned}$$

Does this work? How do we know when to stop? It is worth noting that the residual satisfies  $\mathbf{r} = 0$  if and only if  $\mathbf{x}$  is a stationary point, in our case, a minimizer. One criteria for stopping could then be to check whether  $\|\mathbf{r}_k\|_2 \leq \varepsilon$  for some given tolerance  $\varepsilon > 0$ . One potential problem with this criterion is that the function can become *flat* long before reaching a minimum, so an alternative stopping method would be to stop when the difference between two successive points,  $\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2$ , becomes smaller than some  $\varepsilon > 0$ .

**Example 3.1.** We plot the trajectory of gradient descent with the data

$$\mathbf{A} = \begin{pmatrix} 2 & 0 \\ 1 & 3 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{b} = \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix}.$$

As can be seen from the plot, we always move in the direction orthogonal to a level set, and stop at a point where we are tangent to a level set.

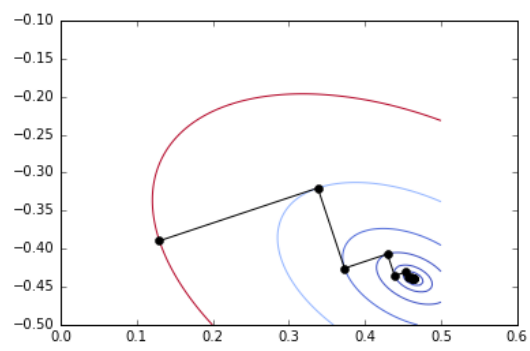


Figure 3.2: Trajectory of gradient descent