Continuous Optimisation, Chpt 2: Unconstrained Optimisation

Peter J.C. Dickinson

DMMP, University of Twente

p.j.c.dickinson@utwente.nl

http://dickinson.website/Teaching/2016CO.html

version: 04/10/16

Monday 19th September 2016

- There is no class next week (Monday 26th September 2016), however the room is available for you.
- Literature: KRT 2.1 and 4.

Table of Contents

Geometry of minimisation

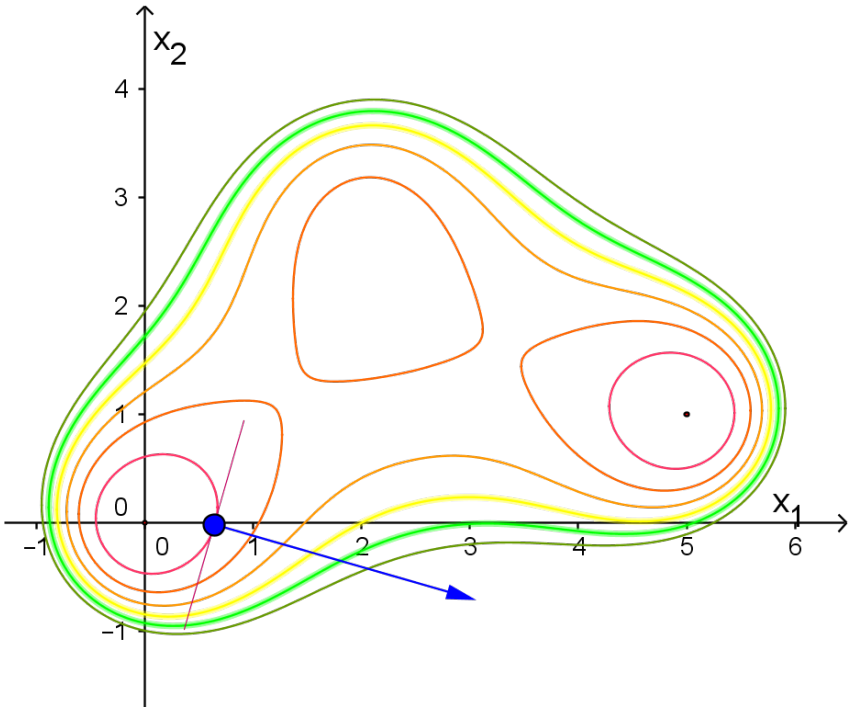
Theorem 2.1 (Geometry of minimisation)

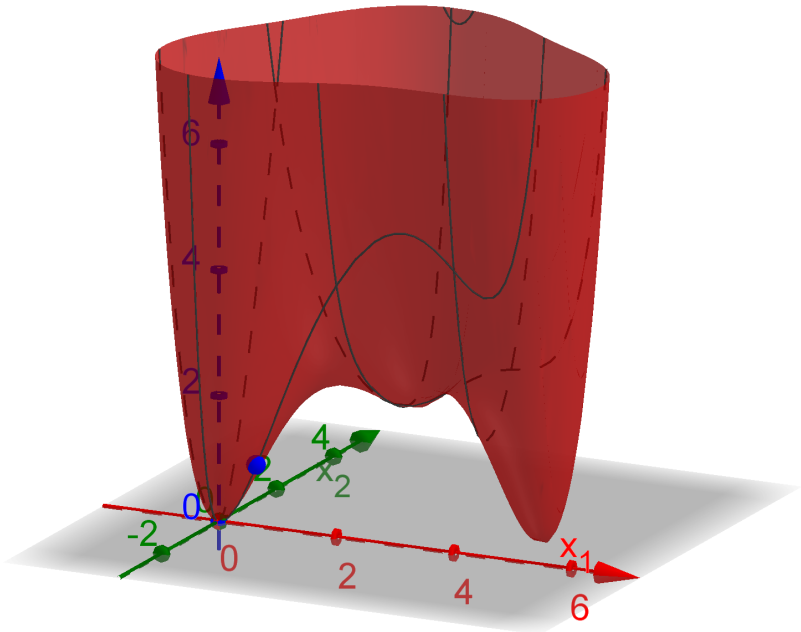
Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathrm{C}^1$ *and a point* $\mathbf{y} \in \mathcal{F}$ *with* $\nabla f(\mathbf{y}) \neq \mathbf{0}$. *In a neighbourhood of* $\mathbf{y}$ *the set* $\mathcal{D}_{\mathbf{y}} = \{\mathbf{x} \in \mathcal{F} : f(\mathbf{x}) = f(\mathbf{y})\}$ *is a* C^1*-manifold of dimension* $n - 1$, *and at* $\mathbf{y}$ *we have* $\nabla f(\mathbf{y}) \perp \mathcal{D}_{\mathbf{y}}$.

Example

http://ggbm.at/e3vayUbW

$$f(x_1, x_2) = \frac{1}{100}(x_1^2 + x_2^2)((x_1 - 5)^2 + (x_2 - 1)^2)((x_1 - 2)^2 + (x_2 - 3)^2 + 1)$$





Two global minima and three strict local minima.

Descent directions

Definition 2.2

For $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$, we call $\mathbf{h} \in \mathbb{R}^n$ a strict descent direction of f at $\mathbf{x}$ if $\exists \bar{\varepsilon} > 0$ s.t. $f(\mathbf{x} + \varepsilon \mathbf{h}) < f(\mathbf{x})$ for all $\varepsilon \in (0, \bar{\varepsilon}]$.

Fill in the quiz at `www.shakeq.com`, login code utwente118.

Descent directions

Definition 2.2

For $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{x} \in \mathbb{R}^n$, we call $\mathbf{h} \in \mathbb{R}^n$ a strict descent direction of f at $\mathbf{x}$ if $\exists \bar{\varepsilon} > 0$ s.t. $f(\mathbf{x} + \varepsilon \mathbf{h}) < f(\mathbf{x})$ for all $\varepsilon \in (0, \bar{\varepsilon}]$.

Fill in the quiz at `www.shakeq.com`, login code utwente118.

Lemma 2.3

For $\mathbf{x} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$ consider the following statements:

1. $\mathbf{x}$ *is a global minimiser of* f*;*
2. $\mathbf{x}$ *is a local minimiser of* f*;*
3. *There are no strict descent directions of* f *at* $\mathbf{x}$*.*

We have $(1) \Rightarrow (2) \Rightarrow (3)$. *If* f *is convex then* $(1) \Leftrightarrow (2) \Leftrightarrow (3)$

Exercises

Ex. 2.1 Prove Lemma 2.3.

Ex. 2.2 Consider $f(x_1, x_2) = (x_1^2 - 2x_2)(x_1^2 - x_2)$. Show that:

(a) the origin $\mathbf{0}$ is not a local minimiser of f;

(b) all $\mathbf{h} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ are strict ascent directions of f at $\mathbf{0}$, i.e. for all $\mathbf{h} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$, $\exists \bar{\varepsilon} > 0$ s.t. $f(\varepsilon \mathbf{h}) > f(\mathbf{0})$ for all $\varepsilon \in (0, \bar{\varepsilon}]$.

N.B. Therefore, in this nonconvex example, statement (3) of Lemma 2.3 holds, but not statement (1).
We thus see that for nonconvex problems even if every direction will lead to an increase, we may still not have a local minimum.

Necessary/sufficient conditions

Theorem 2.4

For $f \in \mathrm{C}^2$ and $\|\mathbf{h}\|$ small:

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^{\mathsf{T}}\mathbf{h} + \tfrac{1}{2}\mathbf{h}^{\mathsf{T}}\nabla^2 f(\mathbf{x})\mathbf{h} + o(\|\mathbf{h}\|^2).$$

Corollary 2.5 (Necessary condition)

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathrm{C}^1$ (resp. $f \in \mathrm{C}^2$). If $\mathbf{x} \in \mathbb{R}^n$ is a local minimiser then $\nabla f(\mathbf{x}) = \mathbf{0}$ (resp. $\nabla^2 f(\mathbf{x}) \succeq 0$).

N.B. Not sufficient, e.g. $f(x) = x^3, \ -x^4$.

Corollary 2.6 (Sufficient condition)

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathrm{C}^2$. If $\mathbf{x} \in \mathbb{R}^n$ has $\nabla f(\mathbf{x}) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}) \succ 0$ then $\mathbf{x}$ is a strict local minimiser of f.

N.B. Not Necessary, e.g. $f(x) = x^4, \ \exp(-x^{-2})$

Convex functions

Corollary 2.7 (from Theorem 1.21 and Corollary 2.5)

For a convex function $f : \mathbb{R}^n \to \mathbb{R}$, $f \in \mathrm{C}^1$ and $\mathbf{x}_0 \in \mathbb{R}^n$ the following are equivalent:

1. $\mathbf{x}_0$ *is a global minimum,*
2. $\mathbf{x}_0$ *is a local minimum,*
3. $\nabla f(\mathbf{x}_0) = \mathbf{0}$.

Lemma 2.8

The set of global minimisers of a convex function is a convex set.

Example: Quadratic functions

For $Q \in \mathcal{S}^n$, $Q \succ 0$, $\mathbf{c} \in \mathbb{R}^n$ and $\gamma \in \mathbb{R}$ consider $f : \mathbb{R}^n \to \mathbb{R}$ given as

$$f(\mathbf{x}) = \mathbf{x}^\mathsf{T} Q\mathbf{x} + 2\mathbf{c}^\mathsf{T}\mathbf{x} + \gamma.$$

We have $\nabla^2 f(\mathbf{x}) = 2Q \succ 0$ and thus f is strictly convex.

A vector $\mathbf{x}$ is a global minimiser of f if and only if

$$\mathbf{0} = \nabla f(\mathbf{x}) = 2Q\mathbf{x} + 2\mathbf{c}.$$

Therefore the unique strict global minimiser is $\mathbf{x}^* = -Q^{-1}\mathbf{c}$, and the optimal value is $f(\mathbf{x}^*) = \gamma - \mathbf{c}^\mathsf{T} Q^{-1}\mathbf{c}$.

Table of Contents

Descent directions and derivatives

Lemma 2.9

Consider $f : \mathbb{R}^n \to \mathbb{R}$, $f \in C^1$, and $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$. Then

- $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{d}}(\mathbf{x}) = \nabla f(\mathbf{x})^\mathsf{T}\mathbf{d}$;
- *If $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{d}}(\mathbf{x}) < 0$ then $\mathbf{d}$ is a strict descent direction of f at $\mathbf{x}$;*
- *If $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{d}}(\mathbf{x}) > 0$ then $\mathbf{d}$ is a strict ascent direction of f at $\mathbf{x}$ (and thus is not a strict descent direction of f at $\mathbf{x}$).*

Basic idea

Basic idea for minimising a function $f : \mathbb{R}^n \to (\mathbb{R} \cup \{\infty\})$, $f \in C^1$ over $\mathbb{R}^n$:

1. Start at a point $\mathbf{x}_0 \in \mathbb{R}^n$. $\quad (k = 0)$
2. Find a **search direction** $\mathbf{d}_k \in \mathbb{R}^n$ such that $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{d}_k}(\mathbf{x}_k) < 0$.
3. If no such direction exists then STOP.
4. **Line search:** Find $\lambda_k = \arg\min_\lambda \{f(\mathbf{x}_k + \lambda \mathbf{d}_k) : \lambda \in \mathbb{R}\}$ (or just $f(\mathbf{x}_k + \lambda_k \mathbf{d}_k) < f(\mathbf{x}_k)$). [See KRT, 4.3]
5. Let $\mathbf{x}_{k+1} = \mathbf{x}_k + \lambda_k \mathbf{d}_k$ and $k \leftarrow k + 1$.
6. If **stopping criteria** satisfied then STOP, else go to step 2.

Choosing $\mathbf{d}$: First order

Lemma 2.10

For $f \in C^1(\mathbb{R}^n, \mathbb{R})$ and $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$ we have $\frac{\partial f}{\partial \mathbf{d}}(\mathbf{x}) = \nabla f(\mathbf{x})^\mathsf{T}\mathbf{d}$ and

$$f(\mathbf{x} + \lambda \mathbf{d}) = f(\mathbf{x}) + \lambda \nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} + o(\lambda).$$

Lemma 2.11

For $f \in C^1(\mathbb{R}^n, \mathbb{R})$ and $\mathbf{x} \in \mathbb{R}^n$ s.t. $\nabla f(\mathbf{x}) \neq \mathbf{0}$ we have

$$\arg\min_{\mathbf{d}} \{\nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} : \|\mathbf{d}\|_2 = 1\} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2}.$$

$\mathbf{d} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|_2}$ *is the direction of steepest descent.*

Ex. 2.3 For $\mathbf{x}_{k+1}, \mathbf{d}_k$ as given on the previous slide with $\lambda_k = \arg\min_\lambda \{f(\mathbf{x}_k + \lambda \mathbf{d}_k) : \lambda \in \mathbb{R}\}$, show that $\mathbf{d}_k^\mathsf{T} \nabla f(\mathbf{x}_{k+1}) = 0$.

Example: Quadratic optimisation

Ex. 2.4 Do exercise 4.15 from KRT.

https://ggbm.at/TYBdQDeB

The convergence to the optimal can be quite slow.

This is a problem in general for minimising a function $f \in C^2$, as if at a minimiser $\mathbf{x}^*$ we have $\nabla^2 f(\mathbf{x}^*) \succ 0$, and then for $A = \nabla^2 f(\mathbf{x}^*) \succ 0, \quad \mathbf{c} = -A\mathbf{x}^*, \quad \gamma = f(\mathbf{x}^*) + \mathbf{x}^{*\mathsf{T}} A\mathbf{x}^*$ we have

$$f(\mathbf{x}) \approx f(\mathbf{x}^*) + (\mathbf{x} - \mathbf{x}^*)^\mathsf{T} A(\mathbf{x} - \mathbf{x}^*) = \mathbf{x}^\mathsf{T} A\mathbf{x} + 2\mathbf{c}^\mathsf{T}\mathbf{x} + \gamma \text{ for } \mathbf{x} \approx \mathbf{x}^*.$$

Newton's method

Lemma 2.12

For $f \in C^1(\mathbb{R}^n, \mathbb{R})$ and $\mathbf{x}, \mathbf{d} \in \mathbb{R}^n$ we have

$$f(\mathbf{x}+\mathbf{d}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\mathsf{T}\mathbf{d} + \tfrac{1}{2}\mathbf{d}^\mathsf{T}\nabla^2 f(\mathbf{x})\mathbf{d} + o(\|\mathbf{d}\|^2).$$

Letting $Q = \frac{1}{2}\nabla^2 f(\mathbf{x}_k)$, $\mathbf{c} = \frac{1}{2}\nabla f(\mathbf{x}_k)$ and $\gamma = f(\mathbf{x}_k)$ we have

$$f(\mathbf{x}_k+\mathbf{d}) \approx \mathbf{d}^\mathsf{T}Q\mathbf{d} + 2\mathbf{c}^\mathsf{T}\mathbf{d} + \gamma.$$

If $Q \succ 0$ then, as a function of $\mathbf{d}$, we have that the right-hand side is minimised at $\mathbf{d} = -Q^{-1}\mathbf{c} = -\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1}\nabla f(\mathbf{x}_k)$.
This is referred to as **Newton's direction**, and in practice works well as a search direction (often with $\lambda_k = 1$).
Finds minimum in one step for quadratic functions.

Ex. 2.5 Show that if $A \succ 0$, $\nabla f(\mathbf{x}_k) \neq \mathbf{0}$ and $\mathbf{d}_k = -A\nabla f(\mathbf{x}_k)$ then $\mathbf{d}_k^\mathsf{T}\nabla f(\mathbf{x}_k) < 0$, and thus $\mathbf{d}_k$ is a descent direction.
Which choices of A give the steepest descent and the Newton's direction respectively?

(Dis)advantages

(+) Newton's method normally converges quicker (in terms of number of steps).

(+) With Newton's method it is normally sufficient to consider a step length of one.

(–) For steepest descent method we need only compute the gradient vector $\nabla f(\mathbf{x}_k)$, whereas for Newton's method we need to also compute the Hessian and $\left(\nabla^2 f(\mathbf{x}_k)\right)^{-1} \nabla f(\mathbf{x}_k)$.

(–) With Newton's method, we require that the Hessian matrix is positive definite.

Exercise

Ex. 2.6 Consider the (convex) function $f : \mathbb{R}^2 \to \mathbb{R}$ given by

$$f(\mathbf{x}) = \exp\left(x_1^2 + 2x_2^2\right).$$

(N.B. The global minimiser is at $\mathbf{x}^* = \mathbf{0}$.)

For the starting point $\mathbf{x}_0 = \begin{pmatrix} 0.6 & 0.6 \end{pmatrix}^\mathsf{T}$, perform the first 7 iterations (i.e. find $\mathbf{x}_1, \ldots, \mathbf{x}_7$) for:

1. the steepest descent method;
2. Newton's method without line search;
3. Newton's method with line search.

Stopping Criteria

Upper bound given by $f(\mathbf{x}_k) \in \mathbb{R}$, i.e. $\inf_{\mathbf{x}}\{f(\mathbf{x}) : \mathbf{x} \in \mathbb{R}^n\} \leq f(\mathbf{x}_k)$.

If also have lower bounds $L_k \in \mathbb{R}$ (e.g. duality, see later in course), can pick parameter $\varepsilon > 0$ and stop when

$$\frac{f(\mathbf{x}_k) - L_k}{1 + |f(\mathbf{x}_k)|} \leq \varepsilon,$$

i.e. relative difference between upper and lower bounds small.

If no (good) lower bounds, can pick parameter $\varepsilon > 0$ and stop when

$$\frac{f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})}{1 + |f(\mathbf{x}_k)|} \leq \varepsilon,$$

i.e. relative improvement small.

Ex. 2.7 Assuming

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \mathbf{d}_k, \qquad \text{and}$$
$$f(\mathbf{x}_{k+1}) \approx f(\mathbf{x}_k) + \mathbf{d}_k^{\mathsf{T}} \nabla f(\mathbf{x}_k) + \tfrac{1}{2} \mathbf{d}_k^{\mathsf{T}} \nabla^2 f(\mathbf{x}_k) \mathbf{d}_k$$

when considering Newton's method, what is

$$f(\mathbf{x}_k) - f(\mathbf{x}_{k+1})$$

approximately equal to?