Theory sheet 10

Second order differential

Last time we have discussed how to approximate $f(\boldsymbol{x}+\Delta\boldsymbol{x})$ for small $\Delta\boldsymbol{x}$ using first order differential:

$$f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + df(\boldsymbol{x},\Delta\boldsymbol{x}),$$

where

$$df(\boldsymbol{x},\Delta\boldsymbol{x}) = \sum_{i=1}^{n} \frac{\partial f}{x_j}\,\Delta x_j = (\operatorname{grad} f(\boldsymbol{x}))^{\top}\Delta\boldsymbol{x}$$

Recall that if we take a step $\boldsymbol{x} \rightsquigarrow \Delta\boldsymbol{x}$ in the direction orthogonal to $\operatorname{grad} f(\boldsymbol{x})$, the above approximation degenerates:

$$f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}),$$

which is trivial. If we still want to catch the effect of this shift on f, we need more precise information, provided by the following object:

Definition 1. *The <u>second order differential</u> of a function $f : \mathbb{R}^n \to \mathbb{R}$ is the following <u>formal</u> object:*

$$d^2 f(\boldsymbol{x}) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i\,\partial x_j}\,dx_i\,dx_j.$$

As with the first order differential, you should think of $d^2 f$ as a function of $\boldsymbol{x}$ and of formal increments dx_j, $j = 1,\ldots,n$. We can then <u>evaluate</u> $d^2 f(\boldsymbol{x})$ on any vector of increments $\Delta\boldsymbol{x}$ as follows:

$$d^2 f(\boldsymbol{x},\Delta\boldsymbol{x}) = \sum_{i,j=1}^{n} \frac{\partial^2 f}{\partial x_i\,\partial x_j}\,\Delta x_i\,\Delta x_j.$$

In the $n = 2$ case it looks as follows:

$$d^2 f(\boldsymbol{x},\Delta\boldsymbol{x}) = f''_{x_1 x_2}(\boldsymbol{x})\,(\Delta x_1)^2 + 2 f''_{x_1 x_2}(\boldsymbol{x})\,\Delta x_1\,\Delta x_2 + f''_{x_2 x_2}(\boldsymbol{x})\,(\Delta x_2)^2.$$

If $\Delta\boldsymbol{x}$ is small, we have the following approximation:

$$\boxed{f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + df(\boldsymbol{x},\Delta\boldsymbol{x}) + \frac{1}{2}\,d^2 f(\boldsymbol{x},\Delta\boldsymbol{x})}$$

If $n = 2$, this reads:

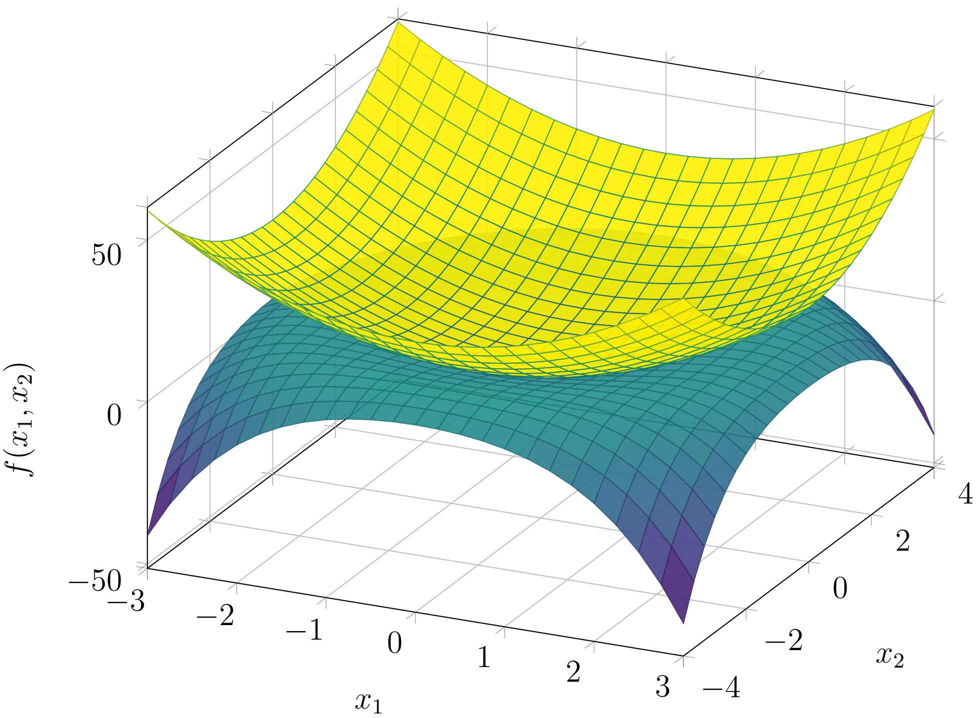
$$f(\boldsymbol{x}+\Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + f'_{x_1}(\boldsymbol{x})\,\Delta x_1 + f'_{x_2}(\boldsymbol{x})\,\Delta x_2 \\ + \frac{1}{2}\,f''_{x_1}(\boldsymbol{x})\,(\Delta x_1)^2 + f''_{x_1 x_2}(\boldsymbol{x})\,\Delta x_1\,\Delta x_2 + \frac{1}{2}\,f''_{x_2}(\boldsymbol{x})\,(\Delta x_2)^2$$

If $n = 1$, this looks even simpler:

$$f(x+\Delta) \approx f(x) + f'(x)\,\Delta + \frac{1}{2}\,f''(x)\,\Delta^2.$$

Here are a few remarks:

- If $\Delta \boldsymbol{x}$ is orthogonal to $\operatorname{grad} f(\boldsymbol{x})$, we have $f(\boldsymbol{x}+\Delta \boldsymbol{x}) \approx f(\boldsymbol{x})+\frac{1}{2} d^2 f(\boldsymbol{x}, \Delta \boldsymbol{x})$, so the dependence on $\Delta \boldsymbol{x}$ does not disappear now!
- For other $\Delta \boldsymbol{x}$, the approximation is now more precise. We say that the boxed formula above gives the second order approximation of f near $\boldsymbol{x}$
- Note that as a function of $\Delta \boldsymbol{x}$ the second order differential $d^2 f(\boldsymbol{x}, \Delta \boldsymbol{x})$ is a quadratic form.
- Geometric interpretation: if first order approximation corresponds to finding a best *line* or plane matching the landscape of f at a given point, the second order approximation gives the best paraboloid/hyperboloid approximation of f near some fixed $\boldsymbol{x}$. Here is an illustration of this:



The yellow surface on this plot is given by $z=3x^2+2y^2$. It is an second order approximation of $z=(3x^2+2y^2)(1-x^2/10-y^2/20)$. Note how the two surfaces touch at $x=y=0$, how they remain close if $(x, y) \approx (0,0)$, but they quickly diverge from each other if we move away from $x=y=0$ too far.

Hessian

Definition 2. *Hessian of a function $f: \mathbb{R}^n \to \mathbb{R}$ at point $\boldsymbol{x} \in \mathbb{R}^n$ is a matrix $H(\boldsymbol{x}) \in M_{n,n}$ of second partial derivatives:*

$$(H(\boldsymbol{x}))_{ij} = \frac{\partial^2 f}{\partial x_i \, \partial x_j}.$$

Similarly to how we used gradient to represent the first order differential by

$$df(\boldsymbol{x}) = (\operatorname{grad} f(\boldsymbol{x}))^\top d\boldsymbol{x},$$

we can use Hessian to represent the second order differential:

$$\boxed{d^2 f(\boldsymbol{x}) = (d\boldsymbol{x})^\top H(\boldsymbol{x})\, d\boldsymbol{x}.}$$

This formula makes the fact that $d^2 f(\boldsymbol{x})$ is a quadratic form of $d\boldsymbol{x}$ mentioned above even clearer.

Note that $H(\boldsymbol{x})$ is symmetric, because partial derivatives commute:

$$(H(\boldsymbol{x}))_{ij} = \frac{\partial^2 f}{\partial x_i\, \partial x_j} = \frac{\partial^2 f}{\partial x_j\, \partial x_i} = (H(\boldsymbol{x}))_{ij}.$$

Combinind the description of $df(\boldsymbol{x})$ in terms of $\operatorname{grad} f(\boldsymbol{x})$ and of $d^2 f(\boldsymbol{x})$ in terms of $H(\boldsymbol{x})$, we obtain

$$\boxed{f(\boldsymbol{x} + \Delta\boldsymbol{x}) \approx f(\boldsymbol{x}) + (\operatorname{grad} f(\boldsymbol{x}))^\top \Delta\boldsymbol{x} + \frac{1}{2}(\Delta\boldsymbol{x})^\top H(\boldsymbol{x})\, \Delta\boldsymbol{x}.}$$

This formula is but another representation of second order approximation of f. Such approximations are called Taylor expansions (of first/second order). There are also Taylor expansions of higher oders (using higher derivatives).

Example

Let $n = 2$, consider $f(x, y) = x^y$ defined for $x, y > 0$. Let us find its Taylor expansion near $x = 1$ and $y = 2$:

$$\begin{aligned}
f(1,2) &= 1^2 = 1 \\
f'_x(1,2) &= yx^{y-1}\,|_{x=1,y=2} = 2 \\
f'_y(1,2) &= x^y \ln x\,|_{x=1,y=2} = 0 \\
f''_{xx}(1,2) &= y(y-1)x^{y-2}\,|_{x=1,y=2} = 2 \\
f''_{xy}(1,2) &= 1 \cdot x^{y-1} + yx^{y-1} \ln x\,|_{x=1,y=2} = 1 \\
f''_{yy}(1,2) &= x^y (\ln x)^2\,|_{x=1,y=2} = 0.
\end{aligned}$$

Therefore,

$$\operatorname{grad} f(1,2) = \begin{pmatrix} 2 \\ 0 \end{pmatrix}, \quad H(1,2) = \begin{pmatrix} 2 & 1 \\ 1 & 0 \end{pmatrix}.$$

Therefore, we have the following second order Taylor approximation:

$$f(1 + \Delta x, 2 + \Delta y) \approx 1 + 2 \cdot \Delta x + 0 \cdot \Delta y + \frac{1}{2} \cdot 2(\Delta x)^2 + 1 \cdot \Delta x\, \Delta y + \frac{1}{2} \cdot 0 \cdot (\Delta y)^2.$$

We can also write this as

$$f(x, y) = 1 + 2 \cdot (x-1) + 0 \cdot (y-2) + \frac{1}{2} \cdot 2(x-1)^2 + 1 \cdot (x-1)\,(y-2) + \frac{1}{2} \cdot 0 \cdot (y-2)^2,$$

where $x = 1 + \Delta x \implies \Delta x = x - 1$ and $y = 2 + \Delta y \implies \Delta y = y - 2$.

For example, our approximation gives

$$f(1.01, 2.01) \approx 1.0202,$$

whereas the exact value is

$$f(1.01, 2.01) = 1.020201508\ldots$$

Free (unconstrained) extrema: first order conditions

Definition 3. *A point $\boldsymbol{x}_0 \in \mathbb{R}^n$ is a point of local maximum of a function $f : \mathbb{R}^n \to \mathbb{R}$ if $f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$ for all for all $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}_0$.*

It is a point of local minimum if $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ for all $\boldsymbol{x}$ sufficiently close to $\boldsymbol{x}_0$.

Recall that to find local extrema of a smooth function of one variable $f : \mathbb{R} \to \mathbb{R}$, we first find candidate extrema points, that is, x such that

$$f'(x) = 0.$$

Some of these points may not be extremal (we need to check second order conditions), but if x is a local extrema, then $f'(x) = 0$ (the conditon is necessary).

If $f : \mathbb{R}^n \to \mathbb{R}$ is a function of many variables and $\boldsymbol{x}$ is its minima or maxima, then in particular it is extrema of a function of one variable $g(x_j) = f(\boldsymbol{x})$ (all other variables are fixed except one; make sure that you understand this argument well!). Therefore, $g'(x_j) = 0$, or

$$g'(x_j) = \frac{\partial f}{\partial x_j} = 0.$$

Therefore, all partial derivatives must be zero at an extremal point. We can write this concisely as:

$$\boxed{f'_{x_j}(\boldsymbol{x}) = 0 \quad \text{for all } j = 1, \ldots, n,}$$

or using gradient notation as

$$\boxed{\operatorname{grad} f(\boldsymbol{x}) = \mathbf{0}.}$$

Let us formulate this as a theorem:

Theorem 1 (First order conditions)**.** *If $\boldsymbol{x}$ is a local minimum or local maximum point of $f : \mathbb{R}^n \to \mathbb{R}$, then*

$$\operatorname{grad} f(\boldsymbol{x}) = \mathbf{0}.$$

As in the univariate case, this is a necessary condition for $\boldsymbol{x}$ to be an extrema, *but not sufficient.*

Remark 1. *Consider $f(x) = x^3$. Clearly, $f'(0) = 3x^2 \mid_{x=0} = 0$, but $x = 0$ is neither minimum, nor maximum of f.*

Type of the extremum: second order conditions

If we found $\boldsymbol{x}_0$ such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$, then we have the following approximation of f near this point:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}_0)^\top H(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0).$$

How do we know if $\boldsymbol{x}_0$ is a minimum or maximum? Can it be neither?

Note that $f(\boldsymbol{x}) \geq f(\boldsymbol{x}_0)$ for $\boldsymbol{x} \approx \boldsymbol{x}_0$ if

$$(\boldsymbol{x} - \boldsymbol{x}_0)^\top H(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0) \geq 0.$$

If $\boldsymbol{x}$ is a local minimum, this condition should hold for all $\boldsymbol{x} \approx \boldsymbol{x}_0$, which by definition means that $H(\boldsymbol{x}_0)$ is positive semi-definite.

Similarly, if $\boldsymbol{x}$ is a point of local maximum, then $f(\boldsymbol{x}) \leq f(\boldsymbol{x}_0)$ for all $\boldsymbol{x} \approx \boldsymbol{x}_0$, which is equivalent to $H(\boldsymbol{x}_0)$ being negative semi-definite.

These are again necessary conditions, but what about sufficient? More precisely, if we found $\boldsymbol{x}_0$ such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$ and $H(\boldsymbol{x}_0)$ is positive semidefinite, can we conclude that f has minimum at this point? The answer is: yes, if $H(\boldsymbol{x}_0)$ is positive definite (strictly).

Theorem 2. *If $\boldsymbol{x}_0$ is a point such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$ and $H(\boldsymbol{x}_0)$ is positive definite, then f has a local minimum at this point.*

If $\boldsymbol{x}_0$ is a point such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$ and $H(\boldsymbol{x}_0)$ is negative definite, then f has a local maximum at this point.

If $\boldsymbol{x}_0$ is a point such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$ and $H(\boldsymbol{x}_0)$ is indefinite, then f does has neither local maximum, nor local minimum at this point.

If $\boldsymbol{x}_0$ is a point such that $\operatorname{grad} f(\boldsymbol{x}_0) = \mathbf{0}$ and $H(\boldsymbol{x}_0)$ is semidefinite (positive or negative), then further analysis is required to say if it is maximum, minimum or neither.

Remark 2. *If $H(\boldsymbol{x}_0)$ is only semidefinite, checking whether $\boldsymbol{x}_0$ is a local extremum requires more work. Consider again the remark above: if $f(x) = x^3$, its Hessian is a 1×1 matrix identified with its second derivative: $H(x) = f''(x) = 6x$. At zero, this Hessian is zero: $H(0) = 0$, so both ≥ 0 and ≤ 0. However, $x_0 = 0$ is clearly not a local extremum of f.*

Consider another example: $f(x) = x^4$. In this case $H(0) = 0$, but we do have a local minimum at $x_0 = 0$. These examples show that some more precise expansions are needed to check extremality at points with semidefinite Hessians.

Remark 3. *Recall that to say that $H(\boldsymbol{x}_0)$ has some type (positive (semi-)definite/negative (semi-)definite, indefinite) is the same as to say that the corresponding quadratic form $d^2 f(\boldsymbol{x}_0)$ has this type. Which is why we frequently talk about $d^2 f(\boldsymbol{x}_0)$ being of some type.*

Remark 4. *Geometrically, positive (negative) definite Hessian means that the function f looks like a paraboloid openning upwards (downwards) near $\boldsymbol{x}_0$. If $H(\boldsymbol{x}_0)$ is indefinite, f looks like a hyperboloid near $\boldsymbol{x}_0$. In this case we say that f has a saddle point at $\boldsymbol{x}_0$.*

Example

Let $f(x,y) = xe^{-x^2-y^2}$. Then

$$f'_x = 0 \implies e^{-x^2-y^2} - 2x^2e^{-x^2-y^2} = 0 \implies x = \pm\frac{1}{\sqrt{2}}$$

and similarly

$$f'_y = 0 \implies -2ye^{-x^2-y^2} = 0 \implies y = 0.$$

Therefore, we have two candidate extremum points: $(\frac{1}{\sqrt{2}}, 0)$ and $(-\frac{1}{\sqrt{2}}, 0)$. We need to check their types:

$$H(\boldsymbol{x}) = \begin{pmatrix} f''_{xx} & f''_{xy} \\ f''_{xy} & f''_{yy} \end{pmatrix} = e^{-x^2-y^2} \begin{pmatrix} 2x(2x^2-3) & 2y(2x^2-1) \\ 2y(2x^2-1) & 2x(2y^2-1) \end{pmatrix}.$$

At $(\frac{1}{\sqrt{2}}, 0)$ we have

$$H\left(\frac{1}{\sqrt{2}}, 0\right) = e^{-\frac{1}{2}} \begin{pmatrix} \frac{-4}{\sqrt{2}} & 0 \\ 0 & -\frac{2}{\sqrt{2}} \end{pmatrix}.$$

This matrix is negative definite, so $(\frac{1}{\sqrt{2}}, 0)$ is a local maximum. Next, at $(-\frac{1}{\sqrt{2}}, 0)$ we have

$$H\left(-\frac{1}{\sqrt{2}}, 0\right) = e^{-\frac{1}{2}} \begin{pmatrix} \frac{4}{\sqrt{2}} & 0 \\ 0 & \frac{2}{\sqrt{2}} \end{pmatrix}.$$

Since this matrix is positive definite, $(-\frac{1}{\sqrt{2}}, 0)$ is a local minimum of f.