II. Calculus

II.04. Optima

Lecture based on

https://github.com/gwthomas/math4ml (Garrett Thomas, 2018)

Prof. Dr. Christoph Lippert

Digital Health & Machine Learning

Taylor's theorem has natural generalizations to functions of more than one variable.

Theorem (Taylor's theorem)

Suppose $f : \mathbb{R}^d \to \mathbb{R}$ is continuously differentiable, and let $\mathbf{h} \in \mathbb{R}^d$.
Then there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x}+t\mathbf{h})^\top \mathbf{h}$$

Furthermore, if f is twice continuously differentiable, then

$$\nabla f(\mathbf{x}+\mathbf{h}) = \nabla f(\mathbf{x}) + \int_0^1 \nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}\,\mathrm{d}t$$

and there exists $t \in (0,1)$ such that

$$f(\mathbf{x}+\mathbf{h}) = f(\mathbf{x}) + \nabla f(\mathbf{x})^\top \mathbf{h} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}+t\mathbf{h})\mathbf{h}$$

This theorem is used in proofs about conditions for local minima of unconstrained optimization problems.

Proposition

If $\mathbf{x}^*$ is a local minimum of f and f is continuously differentiable in a neighborhood of $\mathbf{x}^*$, then $\nabla f(\mathbf{x}^*) = \mathbf{0}$.

Proof.

Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla f(\mathbf{x}^*) \neq \mathbf{0}$.
Let $\mathbf{h} = -\nabla f(\mathbf{x}^*)$, noting that by the continuity of ∇f we have

$$\lim_{t \to 0} -\nabla f(\mathbf{x}^* + t\mathbf{h}) = -\nabla f(\mathbf{x}^*) = \mathbf{h}$$

Hence

$$\lim_{t \to 0} \mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) = \mathbf{h}^\top \nabla f(\mathbf{x}^*) = -\|\mathbf{h}\|_2^2 < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla f(\mathbf{x}^* + t\mathbf{h}) < 0$ for all $t \in [0, T]$.
Now we apply Taylor's theorem:
for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + t\mathbf{h}^\top \nabla f(\mathbf{x}^* + t'\mathbf{h}) < f(\mathbf{x}^*)$$

whence it follows that $\mathbf{x}^*$ is not a local minimum, a contradiction.
Hence $\nabla f(\mathbf{x}^*) = \mathbf{0}$. □

The proof shows us why the vanishing gradient is necessary for an extremum:

- if $\nabla f(\mathbf{x})$ is nonzero, there always exists a sufficiently small step $\alpha > 0$ such that $f(\mathbf{x} - \alpha \nabla f(\mathbf{x}))) < f(\mathbf{x})$.
- For this reason, $-\nabla f(\mathbf{x})$ is called a **descent direction**.
- Points where the gradient vanishes are called **stationary points**.

Note that not all stationary points are extrema.

Consider $f : \mathbb{R}^2 \to \mathbb{R}$ given by $f(x, y) = x^2 - y^2$.

- We have $\nabla f(\mathbf{0}) = \mathbf{0}$, but the point $\mathbf{0}$ is the minimum along the line $y = 0$ and the maximum along the line $x = 0$.
- Thus it is neither a local minimum nor a local maximum of f.

Points such as these, where the gradient vanishes but there is no local extremum, are called **saddle points**.

We have seen that first-order information (i.e. the gradient) is insufficient to characterize local minima.

But we can say more with second-order information (i.e. the Hessian).

First we prove a necessary second-order condition for local minima.

Proposition

If $\mathbf{x}^*$ *is a local minimum of* f *and* f *is twice continuously differentiable in a neighborhood of* $\mathbf{x}^*$*, then* $\nabla^2 f(\mathbf{x}^*)$ *is positive semi-definite.*

Proof.

Let $\mathbf{x}^*$ be a local minimum of f, and suppose towards a contradiction that $\nabla^2 f(\mathbf{x}^*)$ is not positive semi-definite.
Let $\mathbf{h}$ be such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$, noting that by the continuity of $\nabla^2 f$ we have

$$\lim_{t \to 0} \nabla^2 f(\mathbf{x}^* + t\mathbf{h}) = \nabla^2 f(\mathbf{x}^*)$$

Hence

$$\lim_{t \to 0} \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} = \mathbf{h}^\top \nabla^2 f(\mathbf{x}^*)\mathbf{h} < 0$$

Thus there exists $T > 0$ such that $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} < 0$ for all $t \in [0, T]$.
Now we apply Taylor's theorem:
for any $t \in (0, T]$, there exists $t' \in (0, t)$ such that

$$f(\mathbf{x}^* + t\mathbf{h}) = f(\mathbf{x}^*) + \underbrace{t\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}t^2 \mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t'\mathbf{h})\mathbf{h} < f(\mathbf{x}^*)$$

where the middle term vanishes because $\nabla f(\mathbf{x}^*) = \mathbf{0}$ by the previous result.
It follows that $\mathbf{x}^*$ is not a local minimum, a contradiction.
Hence $\nabla^2 f(\mathbf{x}^*)$ is positive semi-definite. □

Now we give sufficient conditions for local minima.

Proposition

Suppose f *is twice continuously differentiable with* $\nabla^2 f$ *positive semi-definite in a neighborhood of* $\mathbf{x}^*$*, and that* $\nabla f(\mathbf{x}^*) = \mathbf{0}$.
Then $\mathbf{x}^*$ *is a local minimum of* f*.*
Furthermore if $\nabla^2 f(\mathbf{x}^*)$ *is positive definite, then* $\mathbf{x}^*$ *is a strict local minimum.*

Proof.

Let B be an open ball of radius $r > 0$ centered at $\mathbf{x}^*$ which is contained in the neighborhood. Applying Taylor's theorem, we have that for any $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, there exists $t \in (0,1)$ such that

$$f(\mathbf{x}^* + \mathbf{h}) = f(\mathbf{x}^*) + \underbrace{\mathbf{h}^\top \nabla f(\mathbf{x}^*)}_{0} + \frac{1}{2}\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq f(\mathbf{x}^*)$$

The last inequality holds because $\nabla^2 f(\mathbf{x}^* + t\mathbf{h})$ is positive semi-definite (since $\|t\mathbf{h}\|_2 = t\|\mathbf{h}\|_2 < \|\mathbf{h}\|_2 < r$), so $\mathbf{h}^\top \nabla^2 f(\mathbf{x}^* + t\mathbf{h})\mathbf{h} \geq 0$.
Since $f(\mathbf{x}^*) \leq f(\mathbf{x}^* + \mathbf{h})$ for all directions $\mathbf{h}$ with $\|\mathbf{h}\|_2 < r$, we conclude that $\mathbf{x}^*$ is a local minimum.
Now further suppose that $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite.
Since the Hessian is continuous we can choose another ball B' with radius $r' > 0$ centered at $\mathbf{x}^*$ such that $\nabla^2 f(\mathbf{x})$ is positive definite for all $\mathbf{x} \in B'$.
Then following the same argument as above (except with a strict inequality now since the Hessian is positive definite) we have $f(\mathbf{x}^* + \mathbf{h}) > f(\mathbf{x}^*)$ for all $\mathbf{h}$ with $0 < \|\mathbf{h}\|_2 < r'$.
Hence $\mathbf{x}^*$ is a strict local minimum. □

Note that, perhaps counterintuitively, the conditions $\nabla f(\mathbf{x}^*) = \mathbf{0}$ and $\nabla^2 f(\mathbf{x}^*)$ positive semi-definite are not enough to guarantee a local minimum at $\mathbf{x}^*$!

Consider the function $f(x) = x^3$.

We have $f'(0) = 0$ and $f''(0) = 0$ (so the Hessian, which in this case is the 1×1 matrix $\begin{bmatrix}0\end{bmatrix}$, is positive semi-definite).

But f has a saddle point at $x = 0$.

The function $f(x) = -x^4$ is an even worse offender – it has the same gradient and Hessian at $x = 0$, but $x = 0$ is a strict local maximum for this function!

For these reasons we require that the Hessian remains positive semi-definite as long as we are close to $\mathbf{x}^*$.

Unfortunately, this condition is not practical to check computationally, but in some cases we can verify it analytically (usually by showing that $\nabla^2 f(\mathbf{x})$ is p.s.d. for all $\mathbf{x} \in \mathbb{R}^d$).

Also, if $\nabla^2 f(\mathbf{x}^*)$ is strictly positive definite, the continuity assumption on f implies this condition, so we don't have to worry.