Functions of several variables: unconstrained extrema

Taylor formula in one dimension

Consider $f : \mathbb{R} \to \mathbb{R}$, possessing N continuous derivatives at $x = x_0$. Then

$$f(x_0 + h) = \sum_{k=0}^{N} \frac{f^{(k)}(x_0)}{k!} h^k + o(h^N) \equiv P_N(x_0, h) + o(h^N), \quad \text{with} \quad \lim_{h \to 0} \frac{o(h^N)}{h^N} = 0. \tag{1}$$

Namely, $P_N(x_0, h)$ is the degree N Taylor polynomial of f at $x = x_0$, while the "error term" $f(x_0 + h) - P_N(x_0, h)$ goes to zero faster than h^N as $h \to 0$. Letting $N \to \infty$, one gets the Taylor series for f, which may or may not converge, depending on f, x_0 and h.

$$f(x_0 + h) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!} h^k. \tag{2}$$

Taylor formula in several dimensions

Consider $f : \mathbb{R}^n \to \mathbb{R}$, possessing continuous partial derivatives up to order N at $\boldsymbol{x} = \boldsymbol{x}_0 = (x_{01}, \ldots, x_{0n})$. Fix a unit direction vector $\boldsymbol{d} = (d_1, \ldots, d_n)$, with $\|\boldsymbol{d}\| = \sqrt{\sum_{i=1}^{n} d_i^2} = 1$. Consider a function $g(t) = f(\boldsymbol{x}_0 + t\boldsymbol{d})$ of one variable $t \in (-1, 1)$, with $g(0) = f(\boldsymbol{x}_0)$ (step away from $\boldsymbol{x}_0$ by the amount t in the direction $\boldsymbol{d}$). The function $g(t)$ can be expanded into the Taylor series (1,2) in t, using the fact that by the Chain rule $\dfrac{d}{dt} = d_1 \dfrac{\partial}{\partial x_1} + \ldots + d_n \dfrac{\partial}{\partial x_n} = \boldsymbol{d} \cdot \nabla$, i.e.

$$\begin{array}{rcccccl}
\frac{d}{dt} g(0) & = & \sum_{i=1}^{n} d_i \frac{\partial}{\partial x_i} f(\boldsymbol{x}_0) & = & (\boldsymbol{d} \cdot \nabla) f(\boldsymbol{x}_0); & & \\
\frac{d^2}{dt^2} g(0) & = & \frac{d}{dt} \left((\boldsymbol{d} \cdot \nabla) f(\boldsymbol{x}_0) \right) & = & \sum_{i=1}^{n} d_i \frac{\partial}{\partial x_i} \left(\sum_{j=1}^{n} d_j \frac{\partial}{\partial x_j} f(\boldsymbol{x}_0) \right) & = & (\boldsymbol{d} \cdot \nabla)^2 f(\boldsymbol{x}_0); \\
& \ldots & & & & & \\
\frac{d^k}{dt^k} g(0) & = & \frac{d}{dt} \left(\frac{d^{k-1}}{dt^k} g(0) \right) & = & (\boldsymbol{d} \cdot \nabla)^k f(\boldsymbol{x}_0). & &
\end{array} \tag{3}$$

In other words, $(\boldsymbol{d} \cdot \nabla)^k$ is a notational shortcut for writing sums of partial derivatives of order k. E.g. if the dimension $n = 2$, with $\boldsymbol{d} = (d_1, d_2)$, the second derivative of g

$$\frac{d^2}{dt^2} g = (\boldsymbol{d} \cdot \nabla)^2 f = d_1^2 \frac{\partial^2}{\partial x_1^2} f + 2 d_1 d_2 \frac{\partial^2}{\partial x_1 \partial x_2} f + d_2^2 \frac{\partial^2}{\partial x_2^2} f = \boldsymbol{d}^T H \boldsymbol{d},$$

where (with yet another notation for partial derivatives)

$$H = \begin{bmatrix} f_{x_1 x_1} & f_{x_1 x_2} \\ f_{x_1 x_2} & f_{x_2 x_2} \end{bmatrix}.$$

In $n = 3$ dimensions still $\frac{d^2}{dt^2} g = \boldsymbol{d}^T H \boldsymbol{d}$, with

$$H = \begin{bmatrix} f_{x_1 x_1} & f_{x_1 x_2} & f_{x_1 x_3} \\ f_{x_1 x_2} & f_{x_2 x_2} & f_{x_2 x_3} \\ f_{x_1 x_3} & f_{x_2 x_3} & f_{x_3 x_3} \end{bmatrix}.$$

In general, for any n the second derivative $\frac{d^2}{dt^2}g$ in the same way equals $\boldsymbol{d}^T H \boldsymbol{d}$, where the matrix

$$H = \begin{bmatrix} \frac{\partial^2}{\partial x_1^2} f & \dots & \frac{\partial^2}{\partial x_1 \partial x_n} f \\ & \dots & \\ \frac{\partial^2}{\partial x_1 \partial x_n} f & \dots & \frac{\partial^2}{\partial x_n^2} f \end{bmatrix}$$

(depending on $\boldsymbol{x}$) is a matrix, called the *Hessian matrix* of f. The Hessian matrix, representing the second derivative of the function of several variables f is therefore often denoted as $D^2 f$.

In the same vein, the notation $Df = \nabla f = \mathrm{grad} f = \left(\frac{\partial}{\partial x_1} f, \dots, \frac{\partial}{\partial x_n} f \right)$ stands for the *derivative*, or the *gradient* of f. Note that given a point $\boldsymbol{x} = \boldsymbol{x}_0$, the derivative (gradient) of the scalar function f at this point is a vector, while the second derivative (the Hessian) is a matrix.

Substituting the notations (3) into the Taylor formulae (1,2), letting $t\boldsymbol{d} = \boldsymbol{h}$, one gets their higher-dimensional analogues:

$$\begin{array}{rcl} f(\boldsymbol{x}_0 + \boldsymbol{h}) & = & \sum_{k=0}^{N} \frac{(\boldsymbol{h} \cdot \nabla)^k f(\boldsymbol{x}_0)}{k!} + o(\|\boldsymbol{h}\|^N), \quad \text{with} \quad \lim_{\boldsymbol{h} \to \boldsymbol{0}} \frac{o(\|\boldsymbol{h}\|^N)}{\|\boldsymbol{h}\|^N} = 0, \\ f(\boldsymbol{x}_0 + \boldsymbol{h}) & = & \sum_{k=0}^{\infty} \frac{(\boldsymbol{h} \cdot \nabla)^k f(\boldsymbol{x}_0)}{k!}. \end{array} \tag{4}$$

In the first formula, the main term (the sum) is a polynomial of degree N in $\boldsymbol{h} = (h_1, \dots, h_n)$, the "error term" goes to zero faster than the length $\|\boldsymbol{h}\|^N$ as $\boldsymbol{h} \to \boldsymbol{0}$. Also, the expression $(\boldsymbol{h} \cdot \nabla)^k f(\boldsymbol{x}_0)$ can be written explicitly by the Newton multinomial formula. For instance, for $n = 2$ with $\boldsymbol{h} = (u, v)$ one has

$$(\boldsymbol{h} \cdot \nabla)^k f(\boldsymbol{x}_0) = \sum_{s=0}^{k} \begin{pmatrix} k \\ s \end{pmatrix} \frac{\partial^k}{\partial x_1^s \partial x_2^{k-s}} f(\boldsymbol{x}_0)\, u^k v^{k-s},$$

where $\begin{pmatrix} k \\ s \end{pmatrix}$ are binomial coefficients, the number of combinations "choose s out of k". More generally, for $\boldsymbol{x} \in \mathbb{R}^n$, we have

$$(\boldsymbol{h} \cdot \nabla)^k f(\boldsymbol{x}_0) = \left(\sum_{j=1}^{n} h_j \frac{\partial}{\partial x_j} \right)^k f(\boldsymbol{x}_0) = \sum_{j_1, \dots, j_n = 1}^{k} h_{j_1} \dots h_{j_k} \frac{\partial^k f(\boldsymbol{x}_0)}{\partial x_{j_1} \dots \partial x_{j_n}}.$$

One can also group the n^k terms in the last formula and rewrite it with multinomial, rather than binomial coefficients. **Note**, importantly, that practically speaking, in all the formulae above, expressions like $\frac{\partial^k f(\boldsymbol{x}_0)}{\partial x_{j_1} \dots \partial x_{j_n}}$ mean that first the partial derivative $\frac{\partial^k f(\boldsymbol{x})}{\partial x_{j_1} \dots \partial x_{j_n}}$ is found, and then one lets $\boldsymbol{x} = \boldsymbol{x}_0$.

We shall be interested in two truncations of the Taylor formula (4):

$$\begin{array}{rcll} f(\boldsymbol{x}_0 + \boldsymbol{h}) & \approx & f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0) \cdot \boldsymbol{h} & \text{and} \\ f(\boldsymbol{x}_0 + \boldsymbol{h}) & \approx & f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0) \cdot \boldsymbol{h} + \frac{1}{2} \boldsymbol{h}^T [D^2 f(\boldsymbol{x}_0)] \boldsymbol{h}. & \end{array} \tag{5}$$

The first truncation approximates the surface $y = f(\boldsymbol{x})$ in $\mathbb{R}^{n+1}$ near $\boldsymbol{x} = \boldsymbol{x}_0$ by its tangent plane $y - y_0 = Df(\boldsymbol{x}_0) \cdot (\boldsymbol{x} - \boldsymbol{x}_0)$, passing through the point $(\boldsymbol{x}, y) = (\boldsymbol{x}_0, f(\boldsymbol{x}_0))$. The second one is still more precise, approximating the surface by a paraboloid.

Remark. Strictly speaking, it makes more sense to accept the above formulas (4,5) as a definition of differentiability of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{x}_0$. I.e. $f(\boldsymbol{x})$ is differentiable at $\boldsymbol{x} = \boldsymbol{x}_0$ if and only if for any $\boldsymbol{h} \in \mathbb{R}^n$ one has

$$f(\boldsymbol{x}_0 + \boldsymbol{h}) = f(\boldsymbol{x}_0) + Df(\boldsymbol{x}_0) \cdot \boldsymbol{h} + o(\|\boldsymbol{h}\|),$$

for some vector $Df(\boldsymbol{x}_0) \in \mathbb{R}^n$. The latter vector is called the *derivative* of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{x}_0$; the components of this vector will be partial derivatives of f. The partial derivatives are defined as *directional derivatives* along the coordinate axes: if $\boldsymbol{d}^i$ is a unit vector in the direction of the ith coordinate axis, $i = 1, \ldots, n$, i.e. the components of $\boldsymbol{d}^i$ are all 0, except 1 at the position i, the limit

$$\frac{\partial f}{\partial x_i}(\boldsymbol{x}_0) = \lim_{t \to 0} \frac{f(\boldsymbol{x}_0 + t\boldsymbol{d}^i) - f(\boldsymbol{x}_0)}{t}$$

is called the ith partial derivative of $f(\boldsymbol{x})$ at $\boldsymbol{x} = \boldsymbol{x}_0$, also often denoted as $f_{x_i}(\boldsymbol{x}_0)$. Higher order partial derivatives are defined in the same way. There is a subtlety here that f is N times differentiable at $\boldsymbol{x} = \boldsymbol{x}_0$ if and only if it has all the partial derivatives up to order N, and they are *continuous*, in which case the order of differentiation for mixed partial derivatives does not matter.

Local extrema

1. `Definition:` A point $\boldsymbol{x} = \boldsymbol{x}_0$ is a local minimum [maximum] of $f(\boldsymbol{x})$ if for all $\boldsymbol{x}$ in some neighbourhood of $\boldsymbol{x}_0$, $f(\boldsymbol{x}) \geq [\leq] f(\boldsymbol{x}_0)$. In both cases, it is called a local extremum. In other words, $\boldsymbol{x} = \boldsymbol{x}_0$ is a local minimum [maximum] of $f(\boldsymbol{x})$ iff for *all* the unit direction vectors $\boldsymbol{d} \in \mathbb{R}^n$, a function of one variable $g(t) = f(\boldsymbol{x}_0 + t\boldsymbol{d})$ has a local minimum [maximum] at $t = 0$.

2. `Note:` Essentially, it means that f will increase [decrease] if one makes a small step away from $\boldsymbol{x} = \boldsymbol{x}_0$ in *any feasible* direction $\boldsymbol{d}$. Dealing with unconstrained extrema, *any* direction is feasible. In a constrained case to be studied later, one will be allowed to step away from $\boldsymbol{x}_0$ only in specific directions, tangent to the intersection of the specific surfaces, corresponding to the *equality constraints.*

3. `Proposition:` If $f(\boldsymbol{x})$ is defined and differentiable in the neighbourhood of $\boldsymbol{x} = \boldsymbol{x}_0$ and has an extremum at $\boldsymbol{x}_0$, then $Df(\boldsymbol{x}_0) = \boldsymbol{0}$. Any point $\boldsymbol{x} \in \mathbb{R}^n$, such that $Df(\boldsymbol{x}) = \boldsymbol{0}$ is called a *critical*, or a *stationary point* of f.
   `Proof:` Otherwise, if $Df(\boldsymbol{x}) \neq \boldsymbol{0}$, f would increase if one steps away from $\boldsymbol{x}_0$ in the direction of grad$f$ and decrease in the opposite direction; hence it cannot have an extremum at $\boldsymbol{x}_0$.
   `Note:` $Df(\boldsymbol{x}_0) = \boldsymbol{0}$ means that the plane, tangent to the surface $y = f(\boldsymbol{x})$ in $\mathbb{R}^{n+1}$ at $\boldsymbol{x} = \boldsymbol{x}_0$ is horizontal (y being the vertical direction).
   `Remark:` The other two possibilities are that f is not defined in a *full* neighbourhood of $\boldsymbol{x}_0$, the latter then being a *boundary point* for the domain of f, or $Df(\boldsymbol{x}_0)$ does not exist, in which case $\boldsymbol{x}_0$ is referred to as a *singular point* of f.

4. `Definition:` A critical point $\boldsymbol{x}_0$, which is not a local extremum of f, is called a *saddle point.* Thus, a critical point is either a local extremum or a saddle point. Classification of local extrema can be in most cases fulfilled via the second derivative test, by looking at the quadratic in $\boldsymbol{h}$ term in the second equation of (5).

5. `Proposition:` A critical point $\boldsymbol{x}_0$ is a local minimum [maximum] if the Hessian $D^2 f(\boldsymbol{x}_0)$ at this point is a positive [negative] definite quadratic form matrix. It is a saddle point if the quadratic form with the matrix $D^2 f(\boldsymbol{x}_0)$ is strictly indefinite.
   `Proof:` Follows from the second formula of (5), where $Df(\boldsymbol{x}_0) = \boldsymbol{0}$, and the ensuing definitions

apropos of quadratic forms.

`Note:` This still leaves behind some (rare and more subtle) possibilities, dealing with the case when the Hessian matrix is *singular* at $\boldsymbol{x} = \boldsymbol{x}_0$ (i.e. it has a zero determinant). These possibilities would require a more subtle analysis, such as looking at the Hessian matrix $D^2 f(\boldsymbol{x})$ for all $\boldsymbol{x}$ in some neighbourhood of $\boldsymbol{x} = \boldsymbol{x}_0$.

Sign-definite and indefinite matrices: quadratic forms

Let $\boldsymbol{h} = (u, v)$ (two dimensions). Let a 2×2 symmetric matrix $Q = \begin{bmatrix} A & B \\ B & C \end{bmatrix}$.

A *quadratic form* in two variables u, v is an expression of the form

$$Q(u, v) = Au^2 + 2Buv + Cv^2 = \boldsymbol{h}^T Q \boldsymbol{h}.$$

Let $\boldsymbol{h} = (u, v, w)$ (three dimensions). Let a 3×3 symmetric matrix $Q = \begin{bmatrix} A & B & C \\ B & D & E \\ C & E & F \end{bmatrix}$.

A *quadratic form* in three variables u, v, w is an expression of the form

$$Q(u, v, w) = Au^2 + Dv^2 + Fw^2 + 2Buv + 2Cuw + 2Evw = \boldsymbol{h}^T Q \boldsymbol{h}.$$

1. `Definition:` Let $\boldsymbol{h} = (h_1, \ldots, h_n) \in \mathbb{R}^n$. Let Q be an $n \times n$ symmetric matrix, i.e $Q^T = Q$. A *quadratic form* in n variables $h_1, \ldots, h_n$ is an expression of the form

$$Q(\boldsymbol{h}) = Q(h_1, \ldots, h_n) = \boldsymbol{h}^T Q \boldsymbol{h} = \boldsymbol{h} \cdot Q\boldsymbol{h}.$$

2. `Definition:` A quadratic form $Q(\boldsymbol{h})$, is said to be

   - positive semi-definite if $\boldsymbol{h} \cdot Q\boldsymbol{h} \geq 0,\ \forall \boldsymbol{h} \neq \boldsymbol{0}$;
   - negative semi-definite if $\boldsymbol{h} \cdot Q\boldsymbol{h} \leq 0,\ \forall \boldsymbol{h} \neq \boldsymbol{0}$;
   - positive definite if $\boldsymbol{h} \cdot Q\boldsymbol{h} > 0,\ \forall \boldsymbol{h} \neq \boldsymbol{0}$;
   - negative definite if $\boldsymbol{h} \cdot Q\boldsymbol{h} < 0,\ \forall \boldsymbol{h} \neq \boldsymbol{0}$;
   - strictly indefinite if $\exists\, \boldsymbol{h}^1, \boldsymbol{h}^2 \neq \boldsymbol{0} :\ \boldsymbol{h}^1 \cdot Q\boldsymbol{h}^1 < 0,\ \boldsymbol{h}^2 \cdot Q\boldsymbol{h}^2 > 0.$

   `Note:` The form Q is (semi)negative definite iff the form $-Q$ is (semi)positive definite.

3. `Eigenvalue criterion:` A quadratic form $Q(h_1, \ldots, h_n)$ with the matrix Q is positive [negative] definite iff all the eigenvalues of Q are strictly positive [negative]. It is strictly indefinite, iff there exists one strictly negative and one strictly positive eigenvalue of the matrix Q.

   `Proof:` Any *symmetric* matrix Q has n mutually orthogonal, unit length eigenvectors $\boldsymbol{u}^i$, $i = 1, \ldots, n$, corresponding to the eigenvalues λ_i, which are all real, but not necessarily distinct. Namely, one has $Q\boldsymbol{u}^i = \lambda_i \boldsymbol{u}^i$, and the number of linearly independent eigenvectors, corresponding to the same eigenvalue λ_i is called the latter's (geometric) *multiplicity*. In any case, the eigenvectors $\boldsymbol{u}^i$ can be chosen[1] to form an orthogonal basis in $\mathbb{R}^n$, and any vector $\boldsymbol{h} \in \mathbb{R}^n$ can be expanded in this basis as

[1] These facts come from linear algebra. For instance, how does one prove that a real symmetric matrix has all real eigenvalues and its eigenvectors, corresponding to different eigenvalues are orthogonal? If $*$ denotes complex conjugate, take the formula $Q\boldsymbol{u} = \lambda \boldsymbol{u}$ and dot it with $\boldsymbol{u}^*$. As $\boldsymbol{u} \neq 0$, $\lambda = \frac{\boldsymbol{u}^* \cdot Q\boldsymbol{u}}{\boldsymbol{u}^* \cdot \boldsymbol{u}}$. The denominator is real, and so is the numerator, because $(\boldsymbol{u}^* \cdot Q\boldsymbol{u})^* = \boldsymbol{u} \cdot Q\boldsymbol{u}^* = \boldsymbol{u}^* \cdot Q\boldsymbol{u}$. The first step used the fact that Q is real, the second that it is symmetric. So λ is real, as well as $\boldsymbol{u}$ can be chosen real. Now suppose $\lambda_{1,2}$ are two distinct eigenvalues and $\boldsymbol{u}^{1,2}$ are corresponding eigenvectors. So $A\boldsymbol{u}^1 = \lambda_1 \boldsymbol{u}^1$ and $A\boldsymbol{u}^2 = \lambda_2 \boldsymbol{u}^2$. Take the dot product of the first expression with $\boldsymbol{u}^2$, the second one with $\boldsymbol{u}^1$ and subtract. By symmetry of Q, get $0 = \boldsymbol{u}^2 \cdot Q\boldsymbol{u}^1 - \boldsymbol{u}^1 \cdot Q\boldsymbol{u}^2 = (\lambda_1 - \lambda_2)(\boldsymbol{u}^1 \cdot \boldsymbol{u}^2)$. So $\boldsymbol{u}^1 \cdot \boldsymbol{u}^2 = 0$, as $\lambda_1 \neq \lambda_2$. If some λ is a multiple eigenvalue of multiplicity k, one can also choose k orthogonal eigenvectors corresponding to it, but proving this is harder.

$\boldsymbol{h} = \sum_{i=1}^{n} h_i \boldsymbol{u}^i$, with some coefficients h_i. Then

$$\boldsymbol{h} \cdot Q\boldsymbol{h} = \left(\sum_{i=1}^{n} h_i \boldsymbol{u}^i \right) \cdot Q \left(\sum_{i=1}^{n} h_i \boldsymbol{u}^i \right) = \sum_{i,j=1,\dots,n} \lambda_i h_i h_j [\boldsymbol{u}^i \cdot \boldsymbol{u}^j] = \sum_{i=1}^{n} \lambda_i h_i^2,$$

as on the last step $\boldsymbol{u}^i \cdot \boldsymbol{u}^j = 0$, whenever $i \neq j$, while $\boldsymbol{u}^i \cdot \boldsymbol{u}^i = \|\boldsymbol{u}^i\|^2 = 1$. This proves the statement, as the right hand side is positive [negative] for all $\boldsymbol{h}$ iff all λ_i are positive [negative], while in the strictly indefinite case one can take $\boldsymbol{h}^1 = \boldsymbol{u}^1$, $\boldsymbol{h}^2 = \boldsymbol{u}^2$, where the corresponding eigenvalues $\lambda_1 < 0$ and $\lambda_2 > 0$.

`Note:` Any quadratic form matrix Q has two important coordinate-independent invariants: its *determinant* $\det Q = \prod_{i=1}^{n} \lambda_i$ and *trace* $\operatorname{Tr} Q = \sum_{i=1}^{n} \lambda_i = \sum_{i=1}^{n} q_{ii}$, the sum of all the *diagonal* elements of Q. Calculating the determinant and the trace can be very informative apropos of the above eigenvalue criterion and fully does the job if the dimension $n = 2$.

`Note:` If $\det Q = 0$, then at least one of its eigenvalues is zero, and in this case the second derivative test fails to determine the type of the critical point. Otherwise, instead of finding all the eigenvalues of Q, one can use the ensuing criterion, due to Sylvestre.

4. `Definition:` Given an $n \times n$ matrix Q, a *leading minor*M_k of Q, of order $k = 1, \dots, n$ is the determinant of the upper left $k \times k$ submatrix of Q.

   `Sylvestre criteiron:` A quadratic form $Q(h_1, \dots, h_n)$ with the matrix Q, such that $\det Q \neq 0$, is

   - positive definite iff $\det M_k > 0$ for all the leading minors M_k of Q, $k = 1, \dots, n$;
   - negative definite iff $(-1)^k \det M_k > 0$ for all the leading minors M_k of Q, $k = 1, \dots, n$;
   - strictly indefinite if none of the above holds.

   `Note:` The proof of the first bullet can be found in a linear algebra textbook. The second bullet easily follows from the first one: If Q is negative-definite, then $-Q$ should be positive definite. However, all the even order leading minors of $-Q$ (corresponding to $k = 2, 4, \dots$) coincide with those for Q, because they are computed in terms of products of k entries of Q, and thus all the $-$ signs will multiply up to 1 rather than -1 for the odd order leading minors (corresponding to $k = 1, 3, \dots$).

   The third bullet also follows easily, because if $\det Q \neq 0$, all the eigenvalues are nonzero, and then the only alternative for then, rather than being all positive [negative] is the existence of a pair of eigenvalues with opposite signs.

   `Note:` The Sylvestre criterion works *only* if $\det Q \neq 0$. For two dimensions (above), positive-definiteness of Q means A, $AC - B^2 > 0$ (and $-A$, $AC - B^2 > 0$ for negative-definiteness). If $AC - B^2 < 0$, the quadratic form is strictly indefinite.