# Meaning of the Hessian of a function in a critical point

Mircea Petrache

February 1, 2012

We consider a function $f : \mathbb{R}^n \to \mathbb{R}$ and assume for it to be differentiable with continuity at least two times (that is, all of the partial derivative functions, which in different notations are written as: $f_{ij} = f_{x_i x_j} = \partial^2_{ij} f = \partial^2_{x_i x_j} f := \frac{\partial^2 f}{\partial x_i \partial x_j} : \mathbb{R}^n \to \mathbb{R}$ for $i, j \in \{1, \dots, n\}$, are continuous).

We recall that a *critical point* of $f$ is by definition a point $P \in \mathbb{R}^n$ such that the gradient of $f$ is zero in $P$:

$$\nabla f(P) = 0, \text{ or, in a more explicit notation, } \begin{pmatrix} \frac{\partial f}{\partial x_1}(P) \\ \vdots \\ \frac{\partial f}{\partial x_n}(P) \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

We are going to see a way (which is actually the most standard one) in which we can (sometimes) tell, for a critical point $P$, whether it is a local maximum, a local minimum, or a saddle point. We will use for that the matrix of second derivatives of $f$, also called the *Hessian matrix* of $f$ at point $P$. This matrix will be denoted by $Hf(P)$ and (after giving a proper meaning to the words "best approximant") is equal to:

$$Hf(P) := \begin{pmatrix} \frac{\partial^2 f}{\partial_{x_1}\partial_{x_1}}(P) & \frac{\partial^2 f}{\partial_{x_1}\partial_{x_2}}(P) & \cdots & \frac{\partial^2 f}{\partial_{x_1}\partial_{x_n}}(P) \\ \frac{\partial^2 f}{\partial_{x_2}\partial_{x_1}}(P) & \frac{\partial^2 f}{\partial_{x_2}\partial_{x_2}}(P) & \cdots & \frac{\partial^2 f}{\partial_{x_2}\partial_{x_n}}(P) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial_{x_n}\partial_{x_1}}(P) & \frac{\partial^2 f}{\partial_{x_n}\partial_{x_2}}(P) & \cdots & \frac{\partial^2 f}{\partial_{x_n}\partial_{x_n}}(P) \end{pmatrix}.$$

# 1 Geometrical meaning of Gradient and Hessian

## 1.1 What does the gradient of a function of one variable represent?

We recall that for a function $f : \mathbb{R} \to \mathbb{R}$ of one single variable, the derivative $f'(P)$ at some point had the following geometrical meaning. Consider the *graph* of $f$, namely the points $G(f) := \{(x, y) \in \mathbb{R}^2 : \ y = f(x)\}$. Usually we draw this subset of $\mathbb{R}^2$ and refer to it directly as "the function $f$", implicitly identifying it to the function itself.
Then recall that $f'(P)$ was representing the *inclination of the tangent line to*

*the graph* $G(f)$ at the point $(P, F(P))$.
In particular, in the case of a *critical point* (in other words, when $f'(P) = 0$) this tangent line at the point where $x = P$ would have to be *horizontal*. That is, the linear function which best approximates $f$ near $P$ is the constant function.

### 1.1.1 Taylor's polynomial of first degree

Another way of finding the lineanction which best approximates $f$ near $P$ is by considering the so-called *Taylor's polyniomial* of first degree. This polynomial, denoted by $j^1_P f(x)$, is defined as:

$$j^1_P f(x) = f(P) + f'(P)(x - P) = (\text{when } f'(P) = 0\ ) = f(P).$$

Another way of thinking about $j^1_P f(x)$ is as the "easiest" function $g$ satisfying $g(P) = f(P)$ and $g'(P) = f'(P)$.

## 1.2 What does the hessian of a function of one variable represent?

Now suppose we want to find how much $f$ deviates from the "best approximant o first degree $j^1_P f$ near $P$". How would you measure this?
We first consider the difference function

$$d(x) := f(x) - j^1_P f(x) = f(x) - f(P) - f'(P)(x - P),$$

and the goal would now be trying to approximate $d$ in the nicest way possible. We observe that now the "best approximant of first degree of $d(x)$" is now just zero, consistently with the idea of a "best approximant": if $d$ would have an approximant $D$ better than zero, then also $j^1_P f + D$ would be a better approximant to $f$ than just $j^1_P f$!
If we want to "follow the trend" established above, the reasonable thing would be to try to find a "best approximant of *second* degree" for $f$. This will be the *second degree* Taylor polynomial, which is (in some sense, which is not made precise here) the "polynomial of second degree which best approximates $f(x)$", and is defined as:

$$j^2_P f(x) = f(P) + f'(P)(x - P) + \frac{1}{2!} f''(P)(x - P)^2.$$

Let's now concentrate on the *new term* which comes in when passing from the best first degree approximant $j^1_P f$ to the best second degree approximant $j^2_P$: this term is

$$\frac{1}{2!} f''(P)(x - P)^2,$$

so it depends basically on the second derivative of $f$ at $P$, i.e on the hessian of $f$!
The meaning of $f''(P)$ is best seen if we consider the case when $P$ is a *critical point* of $f$: as seen above, in this case (i.e. when $f'(P) = 0$), the tangent to the graph of $f$ at the point $(P, f(P))$ is horizontal. We are now saying that $f$ near $P$ is best approximated by the polynomial

$$f(P) + \frac{1}{2!} f''(P)(x - P)^2.$$

This polynomial has as graph a *parabola* with vertex $P$, and with the "arms going up" if $f''(P) > 0$, or with "arms going down" if $f''(P) < 0$. If $f$ stays close fo the parabola with arms going up, then $P$ will be a local minimum for $f$, and if $f$ will stay close to a parabola with arms going down, then $P$ will be a local maximum for $f$. If $f''(P) = 0$ we have that the best approximant of second degree is again not enough in order to really tell the behavior of $f$ near $P$, since as in the case of the best approximant of first degree, this will be again the constant function! To summarize, we can write:

$$\text{If } f'(P) = 0, \text{ then, in case } f''(P) \begin{cases} < 0 & \text{then } P \text{ is a local maximum for } f, \\ > 0 & \text{then } P \text{ is a local minimum for } f, \\ = 0 & \text{then we cannot yet say what kind of critical point } P \text{ is.} \end{cases}$$

Now we can pass to describing the analogous interpretations of gradient vector and hessian matrix in the case when $f$ has more variables, i.e. when $f : \mathbb{R}^n \to \mathbb{R}$ for dimensions $n > 1$.

## 1.3 Gradient and Hessian in more variables

As we saw above, the key to understanding the meaning of the first and second derivatives, at least in the discussion of critical points, is the Taylor's polynomial of $f$. So I would like to write down the formulas for the first and second order Taylor polynomials in the case of a function $f : \mathbb{R}^n \to \mathbb{R}$. The formulas are analogous to the ones for $f : \mathbb{R} \to \mathbb{R}$. We will denote now the coordinates of the point $P \in \mathbb{R}^n$ by $(\bar{x}_1, \ldots, \bar{x}_n) = P$. Then the wanted formula for $j^1_P f$ is (compare with the previous section, and see the similarity!):

$$\begin{aligned} j^1_{(\bar{x}_1,\ldots,\bar{x}_n)} f(x_1, \ldots, x_n) &= f(\bar{x}_1, \ldots, \bar{x}_n) + \\ &\quad + \partial_{x_1} f(\bar{x}_1, \ldots, \bar{x}_n)(x_1 - \bar{x}_1) + \\ &\quad + \partial_{x_2} f(\bar{x}_1, \ldots, \bar{x}_n)(x_2 - \bar{x}_2) + \\ &\quad + \cdots + \\ &\quad + \partial_{x_n} f(\bar{x}_1, \ldots, \bar{x}_n)(x_n - \bar{x}n) \\ &= f(P) + \sum_{i=1}^{n} \partial_{x_i} f(P)(x_i - \bar{x}_i) \\ &= f(P) + \nabla f(P) \cdot (x - P), \end{aligned}$$

where in the last row we denoted by "$\cdot$" the scalar product on vectors of $\mathbb{R}^n$. We recall that if we have two vectors $v = (v_1, \ldots, v_n)$ and $w = (w_1, \ldots, w_n)$ then the definition of "$\cdot$" says that $v \cdot w = v_1 w_1 + \cdots + v_n w_n = \sum_{i=1}^n v_i w_i$. This justifies the last equality in the above formulas for $j^1_P f$.
Again $j^1_P f(x)$ is the polynomial of degree 1 (this time in the variables $x_1, \ldots, x_n$) which best approximates $f$ near $P$.
We can also give (this time with some more immagination) a more "geometric" meaning to the gradient vector $\nabla f(P)$, by considering the graph of $f$: this time the graph is a subset of $\mathbb{R}^{n+1}$, given by all the points $(x, f(x))$ for $x \in \mathbb{R}^n$; if $\nabla f(P) = 0$ then the "tangent hyperplane" to the graph of $f$ will be the one parallel to the horizontal hyperplane $\{(x_1, \ldots, x_n, 0) \in \mathbb{R}^{n+1}\}$. Since in this case $j^1_P f(x) = f(P)$, we see that when approximating $f$ with first degree polynomials near $P$, we cannot distinguish it from a constant function.

Passing to the approximation via second degree polynomials, we obtain the second degree Taylor polynomial

$$\begin{aligned} j^2_P f(x) &= f(P) + \sum_{i=1}^n \partial_{x_i} f(P)(x_i - \bar{x}_i) + \frac{1}{2!} \sum_{i,j=1}^n \partial_{x_i}\partial_{x_j} f(P)(x_i - \bar{x}_i)(x_j - \bar{x}_j) \\ &= f(P) + \nabla f(P) \cdot (x - P) + (x - P)^T \cdot Hf(P) \cdot (x - P). \end{aligned}$$

In the above last row we used the following notation from linear algebra: if $v, w$ are vectors in $\mathbb{R}^n$ and $A$ is a $n \times n$ matrix, then we write

$$v^T \cdot A \cdot w = \sum_{i,j=1}^n a_i A_{ij} w_j.$$

Observe that $Hf(P)$ is the matrix of second derivatives of $f$ at point $P$, having as $(i, j)$-entry the number $\partial_i \partial_j f(P)$, so the last equality of the above formula for $j^2_P f(x)$ is justified.

### 1.3.1 More about the term involving the Hessian matrix, in the case of a critcal point $P$

We will try here to imitate the discussion of Section 1.2 about the meaning of second derivatives of $f$ near a critical point $P$, in the case of $f$ with more than one variable. For a critical point, the function will be well aprroximated by the Taylor polynomial

$$j^2_P f(x) = f(P) + (x - P)^T \cdot Hf(P) \cdot (x - P),$$

and in order to avoid cumbersome notations, we will assume that $P$ is the origin (so that the vector $x - P$ appearing above, which represents the displacement of $x$ from the point $P$, becomes just $= x$). We thus have that our critical point is the origin: $\nabla f(0) = 0$, and so the above formula reads

$$j^2_0 f(x) = f(0) + x^T \cdot Hf(0) \cdot x.$$

In order to continue our discussion, we observe the following property of the matrix of second derivatives of a $C^2$-regular function (i.e. of a function whose second derivatives exist and are continuous):

**Theorem 1** *If $f : \mathbb{R}^n \to \mathbb{R}$ is a function with continuous second derivatives, then the double derivatives commute, i.e. for all $i, j \in \{1, \ldots, n\}$ and for all points $x \in \mathbb{R}^n$ there holds*

$$\frac{\partial^2 f}{\partial x_i \partial x_j}(x) = \frac{\partial^2 f}{\partial x_j \partial x_i}(x).$$

In terms of the Hessian matrix, the above theorem means that $Hf(x)$ is for all $x \in \mathbb{R}^n$ a symmetric matrix (i.e. the $(i, j)$-element of $Hf(x)$ is equal to the $(j, i)$-element: in other words the matrix is "symmetric with respect to the diagonal"). Let's recall more about such matrices.

### 1.3.2 Symmetric matrices (reminders of the Linear Algebra class)

We recall that given a finite dimensional vector space $V$ over the real numbers $\mathbb{R}$ (for example $\mathbb{R}^n$), endowed with a scalar product $\langle \cdot, \cdot \rangle : V \times V \to R$ (for example the euclidean scalar product $\langle v, w \rangle = v \cdot w = \sum_{i=1}^n v_i w_i$ on $\mathbb{R}^n$), a *linear symmetric operator* $L : V \to V$ is a linear operator such that for all vectors $v, w \in V$ there holds $\langle v, L(w) \rangle = \langle L(v), w \rangle$. The following proposition isquite easy to prove:

**Proposition 2** *Suppose $A$ is a $n \times n$ symmetric matrix. Then the function $L_A : \mathbb{R}^n \to \mathbb{R}^n$ defined by*

$$L_A(v) = A \cdot v,$$

*defines a linear symmetric operator, with respect to the euclidean scalar product on $\mathbb{R}^n$ given by $v \cdot W = \sum_{i=1}^n v_i w_i$.*

We recall that a vector $v$ is an *eigenvector* of $L_A$ if $A \cdot v = \lambda v$ for some number $\lambda \in \mathbb{R}$, in other words if $v$ is sent by $L_A$ into a multiple of itself. If $v$ is one of the eigenvectors, then this number $\lambda$ is called an *eigenvalue* (of $A$). (Maybe you know that the eigenvalues for a general matrix $A$, can also be complex numbers. But it is also true that if $A$ is symmetric, then they are actually all *real* numbers!)
You should also remember what an *orthonormal basis* of $\mathbb{R}^n$ is: it is a basis such that all its vectors are of length 1, and perpendicular to each other. The typical example is the usual basis (made of the vectors like $(1, 0, \ldots, 0), (0, 1, 0, \ldots, 0), \ldots$ which have all entries zero but one). In fact *any orthonormal basis can be obtained from the usual one after some rotation* (to be applied to all vectors of the basis) *and/or after changing some vector into its opposite.*
In the next theorem we see that up to the rotations just described above, any linear function $L_A$ coming from a *symmetric matrix* assumes, in the new coordinates, a diagonal form. This is the rigurous statement:

**Theorem 3 (one of the equivalent formulations of the Spectral Theorem)** *Given a symmetric $n \times n$ matrix $A$ with real coefficients, it is possible to find an* orthonormal basis *of* eigenvectors *for the associated linear operator $L_A : \mathbb{R}^n \to \mathbb{R}^n$.*
*In particular, the matrix $B$ representing $L_A$ with respect to this new basis will have the form*

$$\begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix},$$

*where $\lambda_1, \ldots, \lambda_n$ are the* eigenvalues *of $L_A$, repeated according to their multiplicity.*

If you have some practice with linear algebra, you will be able to prove the following lemmas (try first the case of $2 \times 2$ matrices, and then try to generalize!). They are useful if you want to have some idea on the eigenvalues of a matrix $A$.

**Lemma 4** *Let $A$ be a symmetric $n \times n$ matrix. Then the determinant of $A$ is equal to the product of the eigenvalues of the associated linear operator $L_A$ :*

$\mathbb{R}^n \to \mathbb{R}^n$. *[In particular, if $A$ is in diagonal form, then* $\det(A)$ *is the product of the elements on the diagonal.]*

**Lemma 5** *Let $A$ be a symmetric $n \times n$ matrix. Then the trace of $A$ (which is usually defined as the sum of the elements on the diagonal of $A$, i.e. $tr(A) := \sum_{i=1}^{n} a_{ii}$) is equal to the sum of the eigenvalues of $A$.*

# 2 How to apply the knowledge about symmetric matrices to the Taylor polynomial

We saw before that a function $f$ which has "good" (continuous is enough) second derivatives, is best approximated near a critical point (which, say, is the origin 0) by the (second degree) Taylor polynomial

$$j_0^2 F(x) = f(0) + x \cdot Hf(0) \cdot x.$$

now, in order to understand the behavior of $f$ near 0, *we may also "rotate the coordinate basis" like in Theorem 3, so that $Hf(0)$ is in diagonal form, with its eigenvalues on the diagonal!* Why is that more convenient? Because the expression above becomes much simpler: denote indeed by $y = (y_1, \ldots, y_n)$ the coordinates in this new basis of a point $x$ before "rotation". Then (observe that 0 stays 0.. in the general case one would have to rotate the coordiantes around the point $P$) we have

$$\begin{aligned} j_0^2 f(y) &= f(0) + y \cdot [\text{new } Hf(0)] \cdot y \\ &= f(0) + \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}^T \cdot \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix} \cdot \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \\ &= f(0) + \sum_{i=1}^{n} \lambda_i y_i^2. \end{aligned}$$

Now let's interpret the last expression, to see how $j_0^2 f(y)$ behaves for $y$ near 0. *How does this polynomial change if we just vary one coordinate?* Then, since we allow just one coordinate to change, we obtain again a function of only one variable, $y_1$:

$$j_0^2 f(y) = \lambda_1 y_1^2 + f(0) + \sum_{i=2}^{n} \lambda_i y_i^2 = \lambda_1 y_1^2 + constant.$$

The graph of this function can then be drawn in a plane (the horizontal axis will be the $y_1$-axis, the vertical coordinate measures the value of $j_0^2 f$), and it can be discussed like in section 1.2:

1. if $\lambda_1 > 0$ then it is a parabola with "arms up" ,
2. if $\lambda_1 = 0$ then it is a horizontal line, and
3. if $\lambda_1 < 0$ then it is a parabola 'with "arms down".

*Similarly we can vary any coordinate* $y_i$ and the behavior along that direction of $j_0^2 f(y)$ will depend on the value of $\lambda_i$.
*What happens if we change* $y$ *along some direction not parallel to some coordinate axis?* For example, one may want $y$ to change "along the direction" given by the vector $(a_1, \ldots, a_n)$. This would mean that we will take some mumber $t$ close to zero, and consider the multiples of this vector: $y = t \cdot (a_1, \ldots, a_n) = (ta_1, \ldots, ta_n)$. Then we get:

$$\begin{aligned} j_0^2 f(y) &= f(0) + \sum_{i=1}^{n} \lambda_i (ta_i)^2 \\ &= f(0) + t^2 \sum_{i=1}^{n} a_i^2 \lambda_i, \end{aligned}$$

so here we get again a graph shaped like a parabola with "arms up" or "arms down", or a line, depending on whether the coefficient in front of $t^2$ is positive, zero, or negative. *When do these three cases happen?* There is no crystal-clear answer to this. What is clear anyways, is the following:

1. If all the $\lambda_i > 0$, then $\sum_{i=1}^{n} a_i^2 \lambda_i > 0$ for any vector $(a_1, \ldots, a_n)$.
2. If all the $\lambda_i < 0$, then $\sum_{i=1}^{n} a_i^2 \lambda_i < 0$ for any vector $(a_1, \ldots, a_n)$.

In the first case, we can then infer that 0 was a **minimum** of $f$, since $f$ stays close to the Taylor polynomial, which has a minimum, near 0. Similarly, in the second case 0 was a **maximum** of $f$.

*What if the above cases do not happen?* Would this correspond to the case when for functions of one variable the first 2 derivatives were not enough to say if the point was a maximum or a minimum?

Again here the answer is not as simple as expected, since it can happen for functions of more variables that in some directions they have a local minimum and in some others they have a local maximum in the given critical point (this could not happen for functions of 1 variable, where the worst behavior was the presence of a *flex*, as ofr example the point 0 for the function $f(x) = x^3$). This is the case if $Hf(0)$ has some $\lambda_i > 0$ and at the same time some $\lambda_j < 0$: then if we change just $y_i$, $f$ will have a minimum in 0, while if we change just $y_j$ then 0 will look like a maximum in that direction. This kind of critical points are called *saddle points*.

When some $\lambda_i = 0$ then just the second Taylor polynomial is not enough to predict the behavior of $f$ *along that direction*, so one needs to consider better approximants (in this sense the analogy with the 1-dimensional case continues also here).