On the local extrema for functions of several variables

Horia Cornean, 06/11/2012.

1 Some preparatory results

In this section we only work with the Euclidian space $\mathbb{R}^d$, whose norm is defined by $||\mathbf{x}|| = \sqrt{\sum_{j=1}^d |x_j|^2}$. The scalar product between two vectors $\mathbf{x}$ and $\mathbf{y}$ is denoted by $\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{j=1}^d x_j y_j$.

Lemma 1.1. *Let A be a $d \times d$ matrix with real components $\{a_{jk}\}$. Define the quantity $||A||_{\rm HS} := \sqrt{\sum_{j=1}^d \sum_{k=1}^d |a_{jk}|^2}$. Then*

$$||A\mathbf{x}|| \leq ||A||_{\rm HS} \, ||\mathbf{x}||, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{1.1}$$

Proof. From the Cauchy-Schwarz inequality we have:

$$|(A\mathbf{x})_j|^2 = \left(\sum_{k=1}^d a_{jk} x_k \right)^2 \leq \sum_{m=1}^d |a_{jm}|^2 \sum_{n=1}^d |x_n|^2 = \sum_{m=1}^d |a_{jm}|^2 ||\mathbf{x}||^2,$$

and after summation over j we have:

$$||A\mathbf{x}||^2 = \sum_{j=1}^d |(A\mathbf{x})_j|^2 \leq \left(\sum_{j=1}^d \sum_{m=1}^d |a_{jm}|^2 \right) ||\mathbf{x}||^2.$$

□

Lemma 1.2. *Let $K := B_\delta(\mathbf{a}) = \{\mathbf{y} \in \mathbb{R}^d : \, ||\mathbf{y} - \mathbf{a}|| < \delta\}$ be an open ball in $\mathbb{R}^d$. Let $\phi : K \mapsto \mathbb{R}$ be a $C^1(K)$ map (which means that $\partial_j \phi$ exist for all j and are continuous functions on K). Fix $\mathbf{x} \in B_\delta(\mathbf{a})$. Define the real valued function $f(t) = \phi(\mathbf{a} + t(\mathbf{x} - \mathbf{a}))$, $0 \leq t \leq 1$. The function f is continuous on $[0,1]$, differentiable on $(0,1)$, and we have the formula:*

$$f'(t) = \sum_{j=1}^d (x_j - a_j)(\partial_j \phi)(\mathbf{a} + t(\mathbf{x} - \mathbf{a})). \tag{1.2}$$

Proof. Without loss of generality, we assume that $d = 2$. Define $x(t) = a_1 + t(x_1 - a_1)$ and $y(t) = a_2 + t(x_2 - a_2)$. With this notation we have $f(t) = \phi(x(t), y(t))$. Fix $t_0 \in (0,1)$. We may write:

$$\begin{aligned} f(t) - f(t_0) &= \phi(x(t), y(t)) - \phi(x(t_0), y(t_0)) \\ &= \phi(x(t), y(t)) - \phi(x(t_0), y(t)) + \phi(x(t_0), y(t)) - \phi(x(t_0), y(t_0)). \end{aligned} \tag{1.3}$$

For a fixed t, let us define the real valued function $v(s) := \phi(s, y(t))$ on the largest interval which is compatible with the condition that the vector with components $[s, y(t)]$ belongs to K. If $|t - t_0|$ is small enough, then both $x(t)$ and $x(t_0)$ will belong to this interval. We then can apply the mean value theorem for v: there exists some $\tilde{s}$ situated between $x(t_0)$ and $x(t)$ such that

$$v(x(t)) - v(x(t_0)) = v'(\tilde{s})(x(t) - x(t_0)) = (\partial_1 \phi)(\tilde{s}, y(t))(x_1 - a_1)(t - t_0).$$

Thus we constructed some $\tilde{s}$ situated between $x(t_0)$ and $x(t)$ such that

$$\phi(x(t), y(t)) - \phi(x(t_0), y(t)) = (\partial_1\phi)(\tilde{s}, y(t))(x_1 - a_1)(t - t_0).$$

Reasoning in a similar way with the function $v(s) = \phi(x(t_0), s)$, there exists some $\hat{s}$ between $y(t)$ and $y(t_0)$ such that

$$\phi(x(t_0), y(t)) - \phi(x(t_0), y(t_0)) = (\partial_2\phi)(x(t_0), \hat{s})(x_2 - a_2)(t - t_0).$$

Introducing the last two identities in (1.3), if $t \neq t_0$ but $|t - t_0|$ small enough we obtain:

$$\frac{f(t) - f(t_0)}{t - t_0} = (x_1 - a_1)(\partial_1\phi)(\tilde{s}, y(t)) + (x_2 - a_2)(\partial_2\phi)(x(t_0), \hat{s}). \tag{1.4}$$

The distance between the point $[\tilde{s}, y(t)]$ and the point $[x(t_0), y(t_0)]$ tends to zero when t tends to t_0. The same thing happens with the distance between $[x(t_0), \hat{s}]$ and $[x(t_0), y(t_0)]$. Thus the continuity of the partial derivatives of ϕ at $[x(t_0), y(t_0)]$ allows us to write:

$$\begin{aligned} f'(t_0) = \lim_{t \to t_0} \frac{f(t) - f(t_0)}{t - t_0} &= (x_1 - a_1)(\partial_1\phi)(x(t_0), y(t_0)) + (x_2 - a_2)(\partial_2\phi)(x(t_0), y(t_0)) \\ &= \sum_{j=1}^{2} (x_j - a_j)(\partial_j\phi)(\mathbf{a} + t_0(\mathbf{x} - \mathbf{a})). \end{aligned} \tag{1.5}$$

This proves the lemma if $d = 2$. The general case is similar.

□

Lemma 1.3. *Assume that the previous function ϕ is $C^2(K)$ (i.e. the second order partial derivatives exist and are continuous on K). Then $\partial_j\partial_k\phi = \partial_k\partial_j\phi$ on K, for all $1 \leq j, k \leq d$.*

Proof. Without loss of generality, assume that $d = 2$, $j = 1$ and $k = 2$. We will only prove the equality of $\partial_1(\partial_2\phi)(\mathbf{a})$ and $\partial_2(\partial_1\phi)(\mathbf{a})$; the proof is similar for all the other points of K.

If $\mathbf{x}$ is sufficiently close to $\mathbf{a}$, the points with coordinates $[x_1, a_2]$ and $[a_1, x_2]$ belong to K and we can define:

$$g(\mathbf{x}) := \phi(x_1, x_2) - \phi(x_1, a_2) - \phi(a_1, x_2) + \phi(a_1, a_2).$$

Denote by $v(s) = \phi(s, x_2) - \phi(s, a_2)$ the function defined on the maximal interval compatible with the condition that the points $[s, x_2]$ and $[s, a_2]$ belong to K. If $\mathbf{x}$ is sufficiently close to $\mathbf{a}$, then all the real numbers between a_1 and x_1 belong to this interval. We observe that $g(\mathbf{x}) = v(x_1) - v(a_1)$. The mean value theorem applied for v gives us some $\tilde{s}$ between a_1 and x_1 such that:

$$g(\mathbf{x}) = v'(\tilde{s})(x_1 - a_1) = (x_1 - a_1)[(\partial_1\phi)(\tilde{s}, x_2) - (\partial_1\phi)(\tilde{s}, a_2)].$$

Now define the function $u(t) := (\partial_1\phi)(\tilde{s}, t)$ where t varies between a_2 and x_2. We have:

$$g(\mathbf{x}) = (x_1 - a_1)[u(x_2) - u(a_2)] = (x_1 - a_1)(x_2 - a_2)u'(\tilde{t}) = (x_1 - a_1)(x_2 - a_2)\partial_2\partial_1\phi(\tilde{s}, \tilde{t}), \tag{1.6}$$

where $\tilde{t}$ lies between a_2 and x_2.

We will now express g in a different way, using the other mixed second order partial derivative. Define the function $w(t) = \phi(x_1, t) - \phi(a_1, t)$. We have:

$$g(\mathbf{x}) = w(x_2) - w(a_2) = w'(\hat{t})(x_2 - a_2) = (x_2 - a_2)[\partial_2\phi(x_1, \hat{t}) - \partial_2\phi(a_1, \hat{t})]$$

where $\hat{t}$ is between a_2 and x_2. Applying once again the mean value theorem for the function $\partial_2\phi(s, \hat{t})$, we obtain some $\hat{s}$ between a_1 and x_1 such that:

$$g(\mathbf{x}) = (x_1 - a_1)(x_2 - a_2)\partial_1\partial_2\phi(\hat{s}, \hat{t}). \tag{1.7}$$

Comparing (1.6) and (1.7), we see that if $\mathbf{x}$ is close enough to $\mathbf{a}$ but $x_1 \neq a_1$ and $x_2 \neq a_2$, we must have

$$\partial_2\partial_1\phi(\tilde{s},\tilde{t}) = \partial_1\partial_2\phi(\hat{s},\hat{t}),$$

where both points $[\tilde{s},\tilde{t}]$ and $[\hat{s},\hat{t}]$ converge to $\mathbf{a}$ if $||\mathbf{x}-\mathbf{a}||$ converges to zero. The continuity of both partial derivatives at $\mathbf{a}$ finishes the proof. □

If $\phi \in C^2(K)$ and $\mathbf{x} \in K$, we define the Hessian matrix $H(\mathbf{x})$ as the $d \times d$ matrix having the components $H_{jk}(\mathbf{x}) := \partial_j\partial_k\phi(\mathbf{x})$. Because of the previous lemma, we have that the Hessian matrix is self-adjoint.

Lemma 1.4. *Assume that the function ϕ in Lemma 1.1 is $C^2(K)$. Then for every $\mathbf{x} \in K$ there exists some $c_x \in (0,1)$ such that:*

$$\phi(\mathbf{x}) - \phi(\mathbf{a}) = \langle \mathbf{x}-\mathbf{a}, \nabla\phi(\mathbf{a})\rangle + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))(\mathbf{x}-\mathbf{a})\rangle\,. \tag{1.8}$$

Proof. For a fixed j, the function $\partial_j\phi$ is C^1 on K. Define the function $\tilde{f}_j(t) = \partial_j\phi(\mathbf{a}+t(\mathbf{x}-\mathbf{a}))$, where $t \in [0,1]$. The function $\tilde{f}_j$ is differentiable and we can apply formula (1.2) in order to write:

$$\tilde{f}_j'(t) = \sum_{k=1}^{d}(x_k-a_k)\partial_k\partial_j\phi(\mathbf{a}+t(\mathbf{x}-\mathbf{a})).$$

Consider the function $f(t) = \phi(\mathbf{a}+t(\mathbf{x}-\mathbf{a}))$ as in Lemma 1.1. We see from (1.2) that f' is differentiable and we can write:

$$\begin{aligned} f''(t) &= \sum_{j=1}^{d}(x_j-a_j)\tilde{f}_j'(t) = \sum_{j=1}^{d}\sum_{k=1}^{d}(x_j-a_j)(x_k-a_k)\partial_k\partial_j\phi(\mathbf{a}+t(\mathbf{x}-\mathbf{a})) \\ &= \langle \mathbf{x}-\mathbf{a}, H(\mathbf{a}+t(\mathbf{x}-\mathbf{a}))(\mathbf{x}-\mathbf{a})\rangle\,. \end{aligned} \tag{1.9}$$

Moreover, $f'(0) = \sum_{j=1}^{d}(x_j-a_j)\partial_j\phi(\mathbf{a}) = \langle \mathbf{x}-\mathbf{a}, \nabla\phi(\mathbf{a})\rangle$. Now we can apply the Taylor formula with remainder, which provides the existence of some number $c_x \in (0,1)$ such that $f(1)-f(0) = f'(0)+\frac{f''(c_x)}{2}$. The subscript x in the notation of c_x underlines the important fact that this number can change if $\mathbf{x}$ changes. Now since $f(1) = \phi(\mathbf{x})$ and $f(0) = \phi(\mathbf{a})$, the proof is over. □

Lemma 1.5. *Let $\phi \in C^1(K)$. If $\mathbf{a}$ is either a local minimum or maximum, then $\nabla\phi(\mathbf{a}) = 0$.*

Proof. Consider the function $u(t) = \phi(t, a_2, \ldots, a_d)$ defined on the maximal interval $I \subset \mathbb{R}$ which is compatible with the condition that $[t, a_2, \ldots, a_n] \in K$. This interval contains a_1, and a_1 is an interior point of I. Thus a_1 is a local extremum for u, which implies that $u'(a_1) = \partial_1\phi(\mathbf{a}) = 0$. A similar argument shows that all other partial derivatives must be zero at $\mathbf{a}$. □

2 The main results

Theorem 2.1. *Let $\phi \in C^2(K)$ and assume that $\mathbf{a}$ is a critical point (i.e. $\nabla\phi(\mathbf{a}) = 0$). If all the eigenvalues of the Hessian matrix $H(\mathbf{a})$ are positive (negative), then $\mathbf{a}$ is a local minimum (maximum).*

Proof. Using $\nabla\phi(\mathbf{a}) = 0$ in (1.8) we have:

$$\phi(\mathbf{x}) = \phi(\mathbf{a}) + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))(\mathbf{x}-\mathbf{a})\rangle\,. \tag{2.10}$$

Add and substract $\frac{1}{2}\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a})(\mathbf{x}-\mathbf{a})\rangle$ on the right hand side:

$$\phi(\mathbf{x}) = \phi(\mathbf{a}) + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a})(\mathbf{x}-\mathbf{a})\rangle + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))-H(\mathbf{a})](\mathbf{x}-\mathbf{a})\rangle . \quad (2.11)$$

Since $H(\mathbf{a})$ is a self-adjoint matrix, the (complex) spectral theorem insures the existence of an orthonormal basis $\{\Psi_j\}_{j=1}^d$ which consists of eigenvectors of $H(a)$. That is, there exist some real eigenvalues $\{\lambda_j\}_{j=1}^d$ such that $H(\mathbf{a})\Psi_j = \lambda_j\Psi_j$ for all j. Moreover, because all the entries of $H(\mathbf{a})$ are real, the eigenvectors can also be chosen to have real components.

An arbitrary vector $\mathbf{y} \in \mathbb{R}^d$ can be uniquely expressed as $\mathbf{y} = \sum_{j=1}^d \langle \mathbf{y}, \Psi_j\rangle \Psi_j$. Using the linearity of $H(\mathbf{a})$, we have $H(\mathbf{a})\mathbf{y} = \sum_{j=1}^d \langle \mathbf{y}, \Psi_j\rangle H(\mathbf{a})\Psi_j = \sum_{j=1}^d \langle \mathbf{y}, \Psi_j\rangle \lambda_j \Psi_j$. Using the linearity of the scalar product, we have that for every vector $\mathbf{y}$ we can write:

$$\langle \mathbf{y}, H(\mathbf{a})\mathbf{y}\rangle = \sum_{j=1}^d |\langle \mathbf{y}, \Psi_j\rangle|^2 \lambda_j. \quad (2.12)$$

Now assume that all the eigenvalues are positive. Denote by $m > 0$ the smallest of them. Then the above equality becomes:

$$\langle \mathbf{y}, H(\mathbf{a})\mathbf{y}\rangle \geq m \sum_{j=1}^d |\langle \mathbf{y}, \Psi_j\rangle|^2 = m||\mathbf{y}||^2, \quad (2.13)$$

where the last identity is due to the fact that the basis is orthonormal. Replacing $\mathbf{y}$ with $\mathbf{x}-\mathbf{a}$ we have:

$$\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a})(\mathbf{x}-\mathbf{a})\rangle \geq m||\mathbf{x}-\mathbf{a}||^2. \quad (2.14)$$

Introducing this inequality in (2.11) we obtain the inequality:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a}) + \frac{m}{2}||\mathbf{x}-\mathbf{a}||^2 + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))-H(\mathbf{a})](\mathbf{x}-\mathbf{a})\rangle , \quad (2.15)$$

which holds for every $\mathbf{x} \in K$.

Denote by A_x the matrix given by $H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a})) - H(\mathbf{a})$. Using the Cauchy-Schwarz inequality we have:

$$|\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))-H(\mathbf{a})](\mathbf{x}-\mathbf{a})\rangle| = |\langle \mathbf{x}-\mathbf{a}, A_x(\mathbf{x}-\mathbf{a})\rangle| \leq ||\mathbf{x}-\mathbf{a}||\ ||A_x(\mathbf{x}-\mathbf{a})||.$$

Now using Lemma 1.1, we have:

$$|\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a}))-H(\mathbf{a})](\mathbf{x}-\mathbf{a})\rangle| \leq ||\mathbf{x}-\mathbf{a}||^2 ||A_x||_{\mathrm{HS}}.$$

Introducing this in (2.15) we have:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a}) + \frac{1}{2}||\mathbf{x}-\mathbf{a}||^2(m - ||A_x||_{\mathrm{HS}}), \quad (2.16)$$

which holds true on K. Now when $||\mathbf{x}-\mathbf{a}||$ converges to zero, the components a_{jk} of A_x given by

$$a_{jk} = \partial_j\partial_k\phi(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a})) - \partial_j\partial_k\phi(\mathbf{a})$$

will all go to zero independently of the value of $c_x \in (0,1)$ because the second order partial derivatives of ϕ are continuous at $\mathbf{a}$. It means that if $||\mathbf{x}-\mathbf{a}||$ is smaller than some ϵ, then $||A_x||_{\mathrm{HS}}$ can be made smaller than $m/2$. Using this in (2.16), we obtain:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a}) + \frac{m}{4}||\mathbf{x}-\mathbf{a}||^2 \geq \phi(\mathbf{a}), \quad \forall \mathbf{x} \in B_\epsilon(\mathbf{a}) \subset K.$$

This shows that $\mathbf{a}$ is a local minimum for ϕ.

If all the eigenvalues are negative, denote by $-m < 0$ the largest of them. Then (2.12) implies $\langle \mathbf{y}, H(\mathbf{a})\mathbf{y}\rangle \leq -m||\mathbf{y}||^2$ for all $\mathbf{y}$. Using this in (2.11) we obtain:

$$\begin{aligned}\phi(\mathbf{x}) &\leq \phi(\mathbf{a}) - \frac{m}{2}||\mathbf{x}-\mathbf{a}||^2 + \frac{1}{2}\left\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a})) - H(\mathbf{a})](\mathbf{x}-\mathbf{a})\right\rangle \\ &\leq \phi(\mathbf{a}) - \frac{m-||A_x||_{\mathrm{HS}}}{2}||\mathbf{x}-\mathbf{a}||^2,\end{aligned}$$

inequality which holds on K. As before, if ϵ is small enough, then for all $\mathbf{x} \in B_\epsilon(\mathbf{a}) \subset K$ we have that $||A_x||_{\mathrm{HS}} < m/2$ which shows that $\phi(\mathbf{x}) \leq \phi(\mathbf{a})$ on that small ball, hence $\mathbf{a}$ is a local maximum.

□

Theorem 2.2. *Let $\phi \in C^2(K)$ and assume that $\mathbf{a}$ is a critical point (i.e. $\nabla\phi(\mathbf{a}) = 0$). If the Hessian matrix $H(\mathbf{a})$ has at least one positive eigenvalue $\lambda_+ > 0$ and on the same time at least one negative eigenvalue $\lambda_- < 0$, then $\mathbf{a}$ is a saddle point.*

Proof. Denote by $\Psi_\pm$ two real eigenvectors with norm $||\Psi_\pm|| = 1$ corresponding to $\lambda_\pm$. We define the maps $\mathbf{x}_\pm(t) := \mathbf{a} + t\Psi_\pm$ on the maximal intervals $I_\pm \subset \mathbb{R}$ compatible with the condition $\mathbf{x}_\pm(t) \in K$. Clearly, 0 is an interior point for both I_+ and I_-.

Define on I_+ the real valued map $\phi_+(t) := \phi(\mathbf{x}_+(t))$. Replacing $\mathbf{x}$ with $\mathbf{x}_+(t)$ in (2.11) we obtain:

$$\phi_+(t) = \phi(\mathbf{a}) + \frac{\lambda_+ t^2}{2} + \frac{t^2}{2}\left\langle \Psi_+, [H(\mathbf{a}+c_t t\Psi_+) - H(\mathbf{a})]\Psi_+\right\rangle,$$

where the number $c_x \in (0,1)$ got a subscript t in order to explicitly show that it only depends on t. As before, if $|t|$ is smaller than some $\epsilon_+ > 0$, the continuity of the second order partial derivatives of ϕ at $\mathbf{a}$ insure that $||H(\mathbf{a}+c_t t\Psi_+) - H(\mathbf{a})||_{\mathrm{HS}}$ can be made smaller than $\lambda_+/2$. This implies $\phi_+(t) \geq \phi(\mathbf{a}) + \frac{\lambda_+ t^2}{4}$, for all $|t| < \epsilon_+$. In other words, we have constructed points $\mathbf{x} \in K$ which lie arbitrarily close to $\mathbf{a}$ and $\phi(\mathbf{x}) > \phi(\mathbf{a})$.

Now consider $\phi_-(t) = \phi(\mathbf{x}_-(t))$. As above, we obtain:

$$\phi_-(t) = \phi(\mathbf{a}) + \frac{\lambda_- t^2}{2} + \frac{t^2}{2}\left\langle \Psi_-, [H(\mathbf{a}+c_t t\Psi_-) - H(\mathbf{a})]\Psi_-\right\rangle,$$

where again c_t lies somewhere between 0 and 1. Since $|\lambda_-| = -\lambda_- > 0$, there exists $\epsilon_- > 0$ small enough such that if $|t| < \epsilon_-$ we have that $||H(\mathbf{a}+c_t t\Psi_-) - H(\mathbf{a})||_{\mathrm{HS}}$ becomes smaller than $|\lambda_-|/2$. It follows that we have $\phi_-(t) \leq \phi(\mathbf{a}) - \frac{|\lambda_-|t^2}{4}$, for all $|t| < \epsilon_-$. Thus we constructed points $\mathbf{y} \in K$ which lie arbitrary close to $\mathbf{a}$ such that $\phi(\mathbf{y}) < \phi(\mathbf{a})$.

We conclude that $\mathbf{a}$ is a saddle point.

□