On the extrema of functions of several variables

Horia Cornean, 14/08/2014.

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 $u(t) = \phi(t, a_2, \dots, a_d)$ defined on the maximal interval $I \subset \mathbb{R}$ which is compatible with the condition that $[t, a_2, \dots, a_n] \in K$. 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:

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

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 $\{\lambda_j\}_{j=1}^d$ arranged in increasing order 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 $H(\mathbf{a})\mathbf{y} = \sum_{j=1}^d \langle \mathbf{y}, \Psi_j\rangle H(\mathbf{a})\Psi_j = \sum_{j=1}^d \langle \mathbf{y}, \Psi_j\rangle \lambda_j \Psi_j$. Replacing $\mathbf{y}$ with the elements of the standard basis, we can easily obtain the formula:

$$H_{mn}(\mathbf{a}) = \sum_{j=1}^d \lambda_j \Psi_j(m)\Psi_j(n). \tag{2.12}$$

If none of the eigenvalues are zero, we have:

$$[H(\mathbf{a})]^{-1}_{mn} = \sum_{j=1}^d \frac{1}{\lambda_j}\Psi_j(m)\Psi_j(n). \tag{2.13}$$

Using the linearity of the scalar product, we have that for every vector $\mathbf{y}$ we can write:

$$\langle \mathbf{y}, H(\mathbf{a})\mathbf{y}\rangle = \sum_{j=1}^d |\langle \mathbf{y}, \Psi_j\rangle|^2 \lambda_j. \tag{2.14}$$

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

$$\langle \mathbf{y}, H(\mathbf{a})\mathbf{y}\rangle \geq m \sum_{j=1}^d |\langle \mathbf{y}, \Psi_j\rangle|^2 = m||\mathbf{y}||^2, \tag{2.15}$$

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

$$\langle \mathbf{x}-\mathbf{a}, H(\mathbf{a})(\mathbf{x}-\mathbf{a})\rangle \geq m||\mathbf{x}-\mathbf{a}||^2. \tag{2.16}$$

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

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a}) + \frac{m}{2}||\mathbf{x}-\mathbf{a}||^2 + \frac{1}{2}\langle \mathbf{x}-\mathbf{a}, [H(\mathbf{a}+c_x(\mathbf{x}-\mathbf{a})) - H(\mathbf{a})](\mathbf{x}-\mathbf{a})\rangle , \tag{2.17}$$

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 (2.17) we have:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a}) + \frac{1}{2}||\mathbf{x}-\mathbf{a}||^2(m - ||A_x||_{\mathrm{HS}}), \tag{2.18}$$

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 (2.18), 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 (2.14) 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:

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

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:

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

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:

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

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.

□

3 Finding the global minimum of a strictly convex function

In this section we will always assume that the function ϕ is at least C^2 and defined on $\mathbb{R}^d$.

3.1 On convex functions

We say that $\phi : \mathbb{R}^d \mapsto \mathbb{R}$ is convex if:

$$\phi(t\mathbf{x} + (1-t)\mathbf{y}) \le t\phi(\mathbf{x}) + (1-t)\phi(\mathbf{y}), \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \quad 0 < t < 1. \tag{3.19}$$

A function is said to be strictly convex if:

$$\phi(t\mathbf{x} + (1-t)\mathbf{y}) < t\phi(\mathbf{x}) + (1-t)\phi(\mathbf{y}), \quad \forall \mathbf{x} \neq \mathbf{y} \in \mathbb{R}^d, \quad 0 < t < 1. \tag{3.20}$$

Proposition 3.1. *Let $\phi \in C^2(\mathbb{R}^d)$. If the eigenvalues of the Hessian matrix $H(\mathbf{x})$ are always positive, then ϕ is strictly convex.*

Proof. Fix some $\mathbf{x} \neq \mathbf{y}$ and define $g : [0,1] \mapsto \mathbb{R}$, $g(t) = t\phi(\mathbf{x}) + (1-t)\phi(\mathbf{y}) - \phi(t\mathbf{x} + (1-t)\mathbf{y})$. We have $g(0) = g(1) = 0$. We would like to show that $g(t) > 0$ if $0 < t < 1$. This will be achieved in two steps: the first one is to show that g cannot have other zeros between 0 and 1, i.e. it has a constant sign. The second step is to show that this sign is positive.

Assume that g has another zero at $c \in (0,1)$. According to Rolle's theorem, g' should have at least two zeros: one in the interval $(0,c)$ and the other one in the interval $(c,1)$. Let us show that in fact g' can have at most one zero.

We have:

$$g'(t) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(t\mathbf{x} + (1-t)\mathbf{y})\rangle, \quad 0 < t < 1,$$

and

$$g''(t) = -\langle \mathbf{x} - \mathbf{y}, [H(t\mathbf{x} + (1-t)\mathbf{y})](\mathbf{x} - \mathbf{y})\rangle, \quad 0 < t < 1.$$

We see that

$$g'(0) = \phi(\mathbf{x}) - \phi(\mathbf{y}) - \langle \mathbf{x} - \mathbf{y}, \nabla\phi(\mathbf{y})\rangle > 0$$

where the inequality comes from (1.8) and the fact that H has positive eigenvalues. The same fact implies that $g''(t) < 0$, hence g' is strictly decreasing on the interval $(0,1)$. Thus g' can cross zero at most once, therefore g cannot have more than just one zero between 0 and 1.

Now let us prove that g is positive on $(0,1)$. It is enough to do this close to 0, since the sign will be preserved. Using the Taylor formula, there exists some $0 < s < t$ such that:

$$g(t) = g(0) + tg'(0) + \frac{t^2}{2} g''(s)$$

or

$$\frac{g(t)}{t} = g'(0) + \frac{t}{2} g''(s).$$

Since g'' is bounded and because $g'(0) > 0$, the right hand side must be positive for sufficiently small t and we are done. □

Lemma 3.2. *Let $f \in C^2(\mathbb{R})$ be such that $f'' \ge 0$. Then for every $t > 1$ we have:*

$$f(1) - f(0) \le \frac{f(t) - f(1)}{t - 1}.$$

Proof. The mean value theorem provides some $c_1 \in (0,1)$ such that $f(1) - f(0) = f'(c_1)$, and a $c_2 \in (1,t)$ such that $\frac{f(t)-f(1)}{t-1} = f'(c_2)$. Since $f'' \ge 0$ and $c_1 < c_2$, the same mean value theorem implies $f'(c_1) \le f'(c_2)$ and the inequality is proved. □

Lemma 3.3. *Let $\phi \in C^2(\mathbb{R}^d)$ be a strictly convex function with a global minimum. Then ϕ has exactly one critical point $\mathbf{x}_m \in \mathbb{R}^d$, and moreover, $\phi(\mathbf{x}) > \phi(\mathbf{x}_m)$ if $\mathbf{x} \neq \mathbf{x}_m$, i.e. the global minimum of ϕ is only taken in $\mathbf{x}_m$.*

Proof. We note that the assumption about the existence of a global minimum is important. For example, the function e^x is strictly convex on $\mathbb{R}$ but it has no global minimum.

Since ϕ has a global minimum, there must exist some point $\mathbf{x}_m \in \mathbb{R}^d$ such that $\phi(\mathbf{x}) \geq \phi(\mathbf{x}_m)$ for all $\mathbf{x}$. From Lemma 1.5 we know that $\mathbf{x}_m$ is a critical point, i.e. $\nabla\phi(\mathbf{x}_m) = 0$.

Assume that we have another point $\mathbf{x}' \neq \mathbf{x}_m$ such that

$$\phi(\mathbf{x}') = \phi(\mathbf{x}_m) \leq \phi(\mathbf{x}), \quad \forall \mathbf{x} \in \mathbb{R}^d.$$

But since ϕ is strictly convex we would have:

$$\phi(\mathbf{x}'/2 + \mathbf{x}_m/2) < \phi(\mathbf{x}')/2 + \phi(\mathbf{x}_m)/2 = \phi(\mathbf{x}_m)$$

which leads to a contradiction. Thus $\mathbf{x}_m$ is unique. □

Lemma 3.4. *Let $\phi \in C^2(\mathbb{R}^d)$ have a strictly positive definite Hessian matrix on $\mathbb{R}^d$. Assume that there exists a point $\mathbf{a} \in \mathbb{R}^d$ where ϕ takes its global minimum, i.e. $\phi(\mathbf{x}) \geq \phi(\mathbf{a})$ for all $\mathbf{x} \in \mathbb{R}^d$. Choose any $\mathbf{x}_0 \neq \mathbf{a}$. Denote by $I := [\phi(\mathbf{a}), \phi(\mathbf{x}_0)] \subset \mathbb{R}$. Then the set*

$$K := \{\mathbf{x} \in \mathbb{R}^d : \ \phi(\mathbf{a}) \leq \phi(\mathbf{x}) \leq \phi(\mathbf{x}_0)\} = \phi^{-1}(I)$$

is bounded and closed, thus compact.

Proof. Let $\omega \in S^{d-1}$ be an arbitrary element of the unit sphere. The real function $f(t) := \phi(\mathbf{a}+t\omega)$ obeys:

$$f''(t) = \langle \omega, H(\mathbf{a} + t\omega)\omega \rangle > 0, \quad \forall t \in \mathbb{R}.$$

Applying Lemma 3.2 for f we get:

$$\phi(\mathbf{a} + t\omega) \geq \phi(\mathbf{a} + \omega) + (t-1)[\phi(\mathbf{a} + \omega) - \phi(\mathbf{a})], \quad \forall t > 1. \tag{3.21}$$

Because S^{d-1} is compact and ϕ is continuous, the function:

$$S^{d-1} \ni \omega \mapsto \phi(\mathbf{a} + \omega) \in \mathbb{R}$$

is also continuous and attains its minimum at some ω_0. Thus:

$$\phi(\mathbf{a} + \omega) \geq \phi(\mathbf{a} + \omega_0) > \phi(\mathbf{a}), \quad \forall \omega \in S^{d-1}.$$

Using this in (3.21) we have:

$$\phi(\mathbf{a} + t\omega) \geq \phi(\mathbf{a} + \omega_0) + (t-1)[\phi(\mathbf{a} + \omega_0) - \phi(\mathbf{a})], \quad \forall t > 1. \tag{3.22}$$

Now let $\mathbf{x} \notin \overline{B_1(\mathbf{a})}$. Define:

$$\omega := \frac{1}{||\mathbf{x} - \mathbf{a}||}(\mathbf{x} - \mathbf{a}) \in S^{d-1}, \quad t := ||\mathbf{x} - \mathbf{a}|| > 1.$$

We have $\phi(\mathbf{x}) = \phi(\mathbf{a} + t\omega)$ and:

$$\phi(\mathbf{x}) \geq \phi(\mathbf{a} + \omega_0) + (||\mathbf{x} - \mathbf{a}|| - 1)[\phi(\mathbf{a} + \omega_0) - \phi(\mathbf{a})], \quad \mathbf{x} \notin \overline{B_1(\mathbf{a})}.$$

If $||\mathbf{x} - \mathbf{a}||$ is larger or equal than some large enough $R_0 > 1$, then the right hand side of the above inequality can be made larger than $\phi(\mathbf{x}_0)$. Thus no point outside the open ball $B_{R_0}(\mathbf{a})$ can belong to K, which shows that $K \subset B_{R_0}(\mathbf{a})$, hence K is bounded.

Now let us prove that K is also closed. It is enough to prove that it contains all its adherent points. Let $\mathbf{x}$ be such an adherent point; there must exist a sequence $\{\mathbf{x}_n\}_{n\geq 1} \subset K$ such that $\mathbf{x}_n$ converges to $\mathbf{x}$ and

$$\phi(\mathbf{a}) \leq \phi(\mathbf{x}_n) \leq \phi(\mathbf{x}_0), \quad n \geq 1.$$

Since ϕ is continuous, $\phi(\mathbf{x}_n)$ converges to $\phi(\mathbf{x})$. Thus $\phi(\mathbf{a}) \leq \phi(\mathbf{x}) \leq \phi(\mathbf{x}_0)$ and we are done. □

The last technical lemma we need is the following:

Lemma 3.5. *Fix some compact $K \subset \mathbb{R}^d$ and for every $\mathbf{x} \in K$ denote by $\lambda_1(\mathbf{x})$ the smallest eigenvalue of $H(\mathbf{x})$. Then there must exist some number $m > 0$ such that $\lambda_1(\mathbf{x}) \geq m > 0$ for all $\mathbf{x} \in K$.*

Proof. There must exist an eigenvector $\Psi(\mathbf{x}) \in \mathbb{R}^d$ with $||\Psi(\mathbf{x})|| = 1$ and $H(\mathbf{x})\Psi(\mathbf{x}) = \lambda_1(\mathbf{x})\Psi(\mathbf{x})$. Note that $\Psi(\mathbf{x})$ belongs to the unit sphere S^{d-1}.

Since we assumed that $0 < \lambda_1(\mathbf{x})$ for all $\mathbf{x} \in K$ we must have that $0 \leq \inf\{\lambda_1(\mathbf{y}) : \mathbf{y} \in K\}$. If this infimum is positive, then we can choose it as m. Thus the only thing we need to prove is that the infimum cannot equal zero. We will prove this fact by contradiction.

Let us assume that

$$0 = \inf\{\lambda_1(\mathbf{y}) : \mathbf{y} \in K\} < \lambda(\mathbf{x}), \quad \forall \mathbf{x} \in K.$$

This implies that if $n \geq 1$ then $1/n$ cannot be a lower bound for the set $\{\lambda_1(\mathbf{y}) : \mathbf{y} \in K\}$, hence there must exist some $\mathbf{x}_n \in K$ such that $0 < \lambda_1(\mathbf{x}_n) < 1/n$. In this way we constructed a sequence $\{\mathbf{x}_n\}_{n\geq 1} \subset K$ such that $\lim_{n\to\infty} \lambda_1(\mathbf{x}_n) = 0$.

The set S^{d-1} is compact (hence sequentially compact) and the sequence $\{\Psi(\mathbf{x}_n)\}_{n\geq 1}$ belongs to it. Thus one can find a subsequence $\{\Psi(\mathbf{x}_{n_k})\}_{k\geq 1}$ and some unit vector $\Phi \in S^{d-1}$ such that

$$\lim_{k\to\infty} \Psi(\mathbf{x}_{n_k}) = \Phi.$$

Also the (sub)sequence $\{\mathbf{x}_{n_k}\}_{k\geq 1} \subset K$ belongs to a sequentially compact set, thus we may find a (sub)subsequence $\{\mathbf{x}_{n_{k_s}}\}_{s\geq 1} \subset \{\mathbf{x}_{n_k}\}_{k\geq 1}$ and some point $\mathbf{w} \in K$ such that:

$$\lim_{s\to\infty} \mathbf{x}_{n_{k_s}} = \mathbf{w}.$$

To summarize, we have:

$$\lim_{s\to\infty} \lambda_1(\mathbf{x}_{n_{k_s}}) = 0, \quad \lim_{s\to\infty} \mathbf{x}_{n_{k_s}} = \mathbf{w}, \quad \lim_{s\to\infty} \Psi(\mathbf{x}_{n_{k_s}}) = \Phi.$$

Consider the identity:

$$H(\mathbf{w})\Phi = H(\mathbf{w})[\Phi - \Psi(\mathbf{x}_{n_{k_s}})] + [H(\mathbf{w}) - H(\mathbf{x}_{n_{k_s}})]\Psi(\mathbf{x}_{n_{k_s}}) + \lambda_1(\mathbf{x}_{n_{k_s}})\Psi(\mathbf{x}_{n_{k_s}}),$$

which leads to:

$$||H(\mathbf{w})\Phi|| \leq ||H(\mathbf{w})||_{\mathrm{HS}}||\Phi - \Psi(\mathbf{x}_{n_{k_s}})|| + ||H(\mathbf{w}) - H(\mathbf{x}_{n_{k_s}})||_{\mathrm{HS}} + \lambda_1(\mathbf{x}_{n_{k_s}}), \quad \forall s \geq 1. \tag{3.23}$$

Since all the entries of H are continuous at $\mathbf{w}$, we must have

$$\lim_{s\to\infty} ||H(\mathbf{w}) - H(\mathbf{x}_{n_{k_s}})||_{\mathrm{HS}} = 0.$$

Taking $s \to \infty$ in (3.23) leads to $H(\mathbf{w})\Phi = 0$ with $||\Phi|| = 1$. This contradicts the fact that $\lambda_1(\mathbf{w}) > 0$, hence the above infimum must be positive and provides us with the positive lower bound $m > 0$.

□

3.2 The gradient method

Assume that ϕ is as before, and $\mathbf{a}$ is the (unknown) unique critical point which also coincides with the unique point where the global minimum of ϕ is taken. In the following we will explain how one can determine/estimate $\mathbf{a}$ with an arbitrary precision.

Choose any point $\mathbf{x}_0 \in \mathbb{R}^d$. Consider the initial value problem:

$$\mathbf{x}'(t) = -\nabla\phi(\mathbf{x}(t)), \quad \mathbf{x}(0) = \mathbf{x}_0, \ t > 0. \tag{3.24}$$

Since ϕ is a C^2 function, the conditions for the local existence of a solution are satisfied. Moreover, defining $g(t) := ||\nabla\phi(\mathbf{x}(t))||^2 = \sum_{j=1}^{d}[\partial_j\phi(\mathbf{x}(t))]^2$ we have:

$$g'(t) = 2\sum_{j=1}^{d}[\partial_j\phi(\mathbf{x}(t))][\partial_k\partial_j\phi(\mathbf{x}(t))]x_k'(t) = -2\langle\nabla\phi(\mathbf{x}(t)), H(\mathbf{x}(t))\nabla\phi(\mathbf{x}(t))\rangle \leq 0, \quad t > 0.$$

The derivative is non-positive because all the eigenvalues of $H(\mathbf{x}(t))$ are positive, see for comparison (2.16). Thus g is decreasing, which means that $||\nabla\phi(\mathbf{x}(t))||$ becomes smaller and smaller when t grows. Moreover, we can compute:

$$\frac{d}{dt}\phi(\mathbf{x}(t)) = \sum_{j=1}^{d}[\partial_k\phi(\mathbf{x}(t))]x_k'(t) = -||\nabla\phi(\mathbf{x}(t))||^2 \leq 0$$

which shows that the value of $\phi(\mathbf{x}(t))$ decreases with t and must stay trapped in the interval $I = [\phi(\mathbf{a}), \phi(\mathbf{x}_0)]$. We see that both $g'(t)$ and $\frac{d}{dt}\phi(\mathbf{x}(t))$ are zero iff $\nabla\phi(\mathbf{x}(t)) = 0$, otherwise both are negative.

The important extra-information coming from $\phi(\mathbf{a}) \leq \phi(\mathbf{x}(t)) \leq \phi(\mathbf{x}_0)$ is that $\mathbf{x}(t)$ is an element of K. Thus[1] equation (3.24) has a (unique) maximal solution which exists for all $t > 0$. Moreover, Lemma 3.5 implies that there must exist some $m > 0$ such that $\lambda_j(\mathbf{x}) \geq m$ if $\mathbf{x} \in K$. With the same argument as in (2.15) we obtain $g'(t) \leq -2mg(t)$ for all $t > 0$, and:

$$\frac{d}{dt}\{e^{2mt}g(t)\} = 2me^{2mt}g(t) + e^{2mt}g'(t) \leq 0, \quad t > 0$$

which shows that $e^{2mt}g(t)$ is decreasing. In other words:

$$0 \leq g(t) \leq g(0)e^{-2mt}, \quad t \geq 0.$$

Thus $||\nabla\phi(\mathbf{x}(t))||$ goes to zero with t, exponentially fast. This intuitively shows that $\mathbf{x}(t)$ moves towards $\mathbf{a}$, which is the only point where the gradient of ϕ equals zero.

Lemma 3.6. *The solution $\mathbf{x}(t)$ of equation (3.24) converges exponentially fast to $\mathbf{a}$ when $t \to \infty$.*

Proof. Let us first prove that $\mathbf{x}(t)$ has a limit. Let $1 \leq t_1 < t_2$ and use the fundamental theorem of calculus:

$$\mathbf{x}(t_2) - \mathbf{x}(t_1) = \int_{t_1}^{t_2}\mathbf{x}'(t)dt.$$

Then we have:

$$||\mathbf{x}(t_2) - \mathbf{x}(t_1)|| \leq \int_{t_1}^{t_2}||\mathbf{x}'(t)||dt = \int_{t_1}^{t_2}\sqrt{g(t)}dt \leq \frac{\sqrt{g(0)}}{m}(e^{-mt_1} - e^{-mt_2}) \leq \frac{\sqrt{g(0)}}{m}e^{-mt_1}. \quad (3.25)$$

In particular, this shows that the sequence $\{\mathbf{x}(n)\}_{n\geq 1}$ is a Cauchy sequence in K, hence it must have a limit $\mathbf{y} \in K$. Since $||\nabla\phi(\mathbf{x})||$ is continuous we have:

$$0 = \lim_{n\to\infty} g(n) = \lim_{n\to\infty}||\nabla\phi(\mathbf{x}(n))|| = ||\nabla\phi(\mathbf{y})||$$

which shows that $\nabla\phi(\mathbf{y}) = 0$, hence $\mathbf{y} = \mathbf{a}$. Finally, let $t_1 = t$ and $t_2 = n \to \infty$ in (3.25). We have:

$$||\mathbf{a} - \mathbf{x}(t)|| \leq \frac{\sqrt{g(0)}}{m}e^{-mt}, \quad (3.26)$$

which proves the exponentially fast convergence. □

[1]This fact is easier to accept than prove.

If we want to find $\mathbf{a}$ in practice, this method is not always very efficient. Let us from now on assume that we want to determine $\mathbf{a}$ up to a given error $\varepsilon > 0$ while ϕ is regular enough, i.e. at least C^5. From (3.26) we see that we need to estimate $\mathbf{x}(t)$ for a t of order $\ln(1/\varepsilon)$. Now applying a fourth-order Runge-Kutta iteration with step h, the number of iterations being given by $N = t/h$, we can find $\mathbf{x}(t)$ up to an error of order $N\, h^5 = t^5/N^4$. Thus we need to choose

$$N \sim \varepsilon^{-\frac{1}{4}}[\ln(1/\varepsilon)]^{\frac{5}{4}}.$$

Thus if $\varepsilon \sim 10^{-1}$ then $N \sim 5$, if $\varepsilon \sim 10^{-6}$ then $N \sim 850$, and if $\varepsilon \sim 10^{-10}$ then $N \sim 16000$.

3.3 An iterative method for finding the critical point

Now let us show how we can combine the previous method with a second order (super-exponential) iterative method in order to increase computational efficiency. We start with three technical lemmas.

Lemma 3.7. *Let $\phi \in C^3(\mathbb{R}^d)$. There exists a numerical constant $c_1 < \infty$ such that for every $\mathbf{u}, \mathbf{w} \in \overline{B_1(\mathbf{a})}$ we have:*

$$\max\left\{||H(\mathbf{u}) - H(\mathbf{w})||_{\mathrm{HS}},\ ||[H(\mathbf{u})]^{-1} - [H(\mathbf{w})]^{-1}||_{\mathrm{HS}}\right\} \leq c_1\, ||\mathbf{u} - \mathbf{w}||.$$

Proof. Define

$$h_{jk}(s) := \partial_j\partial_k\phi(\mathbf{w} + s(\mathbf{u} - \mathbf{w})), \quad 0 \leq s \leq 1,\ \mathbf{u}, \mathbf{w} \in \overline{B_1(\mathbf{a})}.$$

There exists some $s_{\mathbf{u},\mathbf{w},j,k} \in (0, 1)$ such that $h_{jk}(1) - h_{jk}(0) = h'_{jk}(s_{\mathbf{u},\mathbf{w},j,k})$ or:

$$\partial_j\partial_k\phi(\mathbf{u}) - \partial_j\partial_k\phi(\mathbf{w}) = \sum_{m=1}^{d} \partial_m\partial_j\partial_k\phi(\mathbf{w} + s_{\mathbf{u},\mathbf{w},j,k}(\mathbf{u} - \mathbf{w}))](u_m - w_m). \tag{3.27}$$

In terms of matrix elements:

$$H_{jk}(\mathbf{u}) - H_{jk}(\mathbf{w}) = \sum_{m=1}^{d} \partial_m\partial_j\partial_k\phi(\mathbf{w} + s_{\mathbf{u},\mathbf{w},j,k}(\mathbf{u} - \mathbf{w}))](u_m - w_m). \tag{3.28}$$

The vector $\mathbf{w} + s_{\mathbf{u},\mathbf{w},j,k}(\mathbf{u} - \mathbf{w}))$ always belongs to $\overline{B_1(\mathbf{a})}$. Because $\phi \in C^3(\mathbb{R}^d)$ and $\overline{B_1(\mathbf{a})}$ is compact, we have that

$$s_1 := \max_{m,j,k \in \{1,\ldots,d\}} \sup_{\mathbf{x} \in \overline{B_1(\mathbf{a})}} |\partial_m\partial_j\partial_k\phi(\mathbf{x})| < \infty. \tag{3.29}$$

Thus:

$$|H_{jk}(\mathbf{u}) - H_{jk}(\mathbf{w})| \leq s_1\sqrt{d}||\mathbf{u} - \mathbf{w}||, \quad j, k \in \{1, ..., d\},$$

or

$$||H(\mathbf{u}) - H(\mathbf{w})||_{\mathrm{HS}} \leq s_1 d^{3/2}||\mathbf{u} - \mathbf{w}||,$$

which proves one of the estimates of the lemma. The second one uses the following identity:

$$[H(\mathbf{u})]^{-1} - [H(\mathbf{w})]^{-1} = [H(\mathbf{u})]^{-1}\{H(\mathbf{w}) - H(\mathbf{u})\}[H(\mathbf{w})]^{-1},$$

from which we can bound the norm of the left hand side:

$$||[H(\mathbf{u})]^{-1} - [H(\mathbf{w})]^{-1}||_{\mathrm{HS}} \leq ||[H(\mathbf{u})]^{-1}||_{\mathrm{HS}}\ ||H(\mathbf{u}) - H(\mathbf{w})||_{\mathrm{HS}}\ ||[H(\mathbf{w})]^{-1}||_{\mathrm{HS}}.$$

The entries of both $[H(\mathbf{w})]^{-1}$ and $[H(\mathbf{u})]^{-1}$ are continuous on $\overline{B_1(\mathbf{a})}$, thus their Hilbert-Schmidt norms can be bounded from above by some numerical constant. One can prove a better estimate. If

$$m_1 := \inf_{\mathbf{x} \in \overline{B_1(\mathbf{a})}} \lambda_1(\mathbf{x}) > 0, \quad M_d := \sup_{\mathbf{x} \in \overline{B_1(\mathbf{a})}} \lambda_d(\mathbf{x}) \tag{3.30}$$

then using (2.12) and (2.13) together with the the fact that the eigenvectors Ψ_j form an orthonormal basis, one can prove:

$$||[H(\mathbf{x})]||_{\mathrm{HS}} \leq d^{1/2} M_d, \quad ||[H(\mathbf{x})]^{-1}||_{\mathrm{HS}} \leq \frac{d^{1/2}}{m_1}, \quad \forall \mathbf{x} \in \overline{B_1(\mathbf{a})}. \tag{3.31}$$

Putting everything together, we see that we can choose

$$c_1 = \max\{s_1 d^{3/2},\ s_1 d^{5/2}/m_1^2\} \tag{3.32}$$

and the proof is over. □

Lemma 3.8. *Let $\phi \in C^3(\mathbb{R}^d)$. With the same numerical constant c_1 as in the previous lemma, and for every $\mathbf{y}, \mathbf{z} \in \overline{B_1(\mathbf{a})}$ we have:*

$$||\nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{z}) - [H(\mathbf{y})](\mathbf{y}-\mathbf{z})|| \leq c_1\, ||\mathbf{y}-\mathbf{z}||^2, \quad \mathbf{z}, \mathbf{y} \in \overline{B_1(\mathbf{a})}. \tag{3.33}$$

Proof. Define

$$h_j(t) := \partial_j\phi(\mathbf{z} + t(\mathbf{y}-\mathbf{z})), \quad 0 \leq t \leq 1,\ \mathbf{z}, \mathbf{y} \in \overline{B_1(\mathbf{a})}.$$

There exists some $t_{\mathbf{z},\mathbf{y},j} \in (0,1)$ such that $h_j(1) - h_j(0) = h_j'(t_{\mathbf{z},\mathbf{y},j})$ or:

$$\partial_j\phi(\mathbf{y}) - \partial_j\phi(\mathbf{z}) = \{[H(\mathbf{z} + t_{\mathbf{z},\mathbf{y},j}(\mathbf{y}-\mathbf{z}))](\mathbf{y}-\mathbf{z})\}_j, \tag{3.34}$$

or even more:

$$\partial_j\phi(\mathbf{y}) - \partial_j\phi(\mathbf{z}) = \{[H(\mathbf{y})](\mathbf{y}-\mathbf{z})\}_j + \{[H(\mathbf{z} + t_{\mathbf{z},\mathbf{y},j}(\mathbf{y}-\mathbf{z})) - H(\mathbf{y})](\mathbf{y}-\mathbf{z})\}_j. \tag{3.35}$$

Denote by $\mathbf{u} = \mathbf{z} + t_{\mathbf{z},\mathbf{y},j}(\mathbf{y}-\mathbf{z}) \in \overline{B_1(\mathbf{a})}$ and apply Lemma 3.7 to the pair $\mathbf{u}$ and $\mathbf{w} = \mathbf{y}$. Since $\mathbf{u} - \mathbf{w} = (1 - t_{\mathbf{z},\mathbf{y},j})(\mathbf{y}-\mathbf{z})$, then we have:

$$||[H(\mathbf{u}) - H(\mathbf{w})](\mathbf{y}-\mathbf{z})|| \leq ||[H(\mathbf{u}) - H(\mathbf{w})]||_{\mathrm{HS}}||\mathbf{y}-\mathbf{z}|| \leq c_1\, ||\mathbf{y}-\mathbf{z}||^2$$

and we are done. □

Lemma 3.9. *Let $\phi \in C^3(\mathbb{R}^d)$. For any $0 < \delta < 1$ we define $\mathbf{f}_\delta : \overline{B_\delta(\mathbf{a})} \mapsto \mathbb{R}^d$ given by $\mathbf{f}_\delta(\mathbf{x}) := \mathbf{x} - [H(\mathbf{x})]^{-1}[\nabla\phi(\mathbf{x})]$. Then there exists a numerical constant $C_1 := d^{1/2}c_1/m_1$ such that:*

$$||\mathbf{f}_\delta(\mathbf{x}) - \mathbf{a}|| \leq C_1\, ||\mathbf{x}-\mathbf{a}||^2, \quad \mathbf{x} \in \overline{B_\delta(\mathbf{a})}. \tag{3.36}$$

Moreover, there exists a small enough δ_0 such that $\mathbf{f}_{\delta_0}$ leaves $\overline{B_{\delta_0}(\mathbf{a})}$ invariant and

$$||\mathbf{f}_{\delta_0}(\mathbf{y}) - \mathbf{f}_{\delta_0}(\mathbf{y})|| \leq \frac{1}{2}||\mathbf{y}-\mathbf{z}||,$$

i.e. $\mathbf{f}_{\delta_0}$ is a contraction.

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

$$||\mathbf{f}_\delta(\mathbf{x}) - \mathbf{a}|| = ||\mathbf{x} - \mathbf{a} - [H(\mathbf{x})]^{-1}[\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{a})]|| = ||[H(\mathbf{x})]^{-1}\{\nabla\phi(\mathbf{x}) - \nabla\phi(\mathbf{a}) - [H(\mathbf{x})](\mathbf{x}-\mathbf{a})\}||$$

and using (3.31) together with (3.33) where we put $\mathbf{z} = \mathbf{x}$ and $\mathbf{y} = \mathbf{a}$ we obtain (3.36). It follows that if δ is small enough such that $C_1\delta < 1$, then $\mathbf{f}_\delta(\mathbf{x}) \in \overline{B_\delta(\mathbf{a})}$, which means that $\mathbf{f}_\delta$ leaves $\overline{B_\delta(\mathbf{a})}$ invariant. Moreover, by a simple computation we obtain:

$$\mathbf{f}_\delta(\mathbf{y}) - \mathbf{f}_\delta(\mathbf{z}) = -[H(\mathbf{y})]^{-1}\{\nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{z}) - [H(\mathbf{y})](\mathbf{y}-\mathbf{z})\} + \{[H(\mathbf{z})]^{-1} - [H(\mathbf{y})]^{-1}\}\nabla\phi(\mathbf{z}). \tag{3.37}$$

From (3.31), (3.33) and the definition of C_1 we obtain:

$$||[H(\mathbf{y})]^{-1}\{\nabla\phi(\mathbf{y}) - \nabla\phi(\mathbf{z}) - [H(\mathbf{y})](\mathbf{y}-\mathbf{z})\}|| \leq C_1||\mathbf{y}-\mathbf{z})||^2 \leq C_1\delta\ ||\mathbf{y}-\mathbf{z}||.$$

From Lemma 3.7 we obtain:

$$||\{[H(\mathbf{z})]^{-1} - [H(\mathbf{y})]^{-1}\}\nabla\phi(\mathbf{z})|| \leq c_1||\mathbf{z}-\mathbf{y}||\ ||\nabla\phi(\mathbf{z})|| = c_1||\mathbf{z}-\mathbf{y}||\ ||\nabla\phi(\mathbf{z}) - \nabla\phi(\mathbf{a})||,$$

where we used that $\nabla\phi(\mathbf{a}) = 0$. From (3.34) we obtain:

$$||\nabla\phi(\mathbf{z}) - \nabla\phi(\mathbf{a})|| \leq \sup_{\mathbf{x}\in\overline{B_1(\mathbf{a})}} ||H(\mathbf{x})||_{\mathrm{HS}}\ ||\mathbf{z}-\mathbf{a}|| \leq d^{1/2}M_d\ \delta$$

which gives:

$$||\{[H(\mathbf{z})]^{-1} - [H(\mathbf{y})]^{-1}\}\nabla\phi(\mathbf{z})|| \leq c_1\ d^{1/2}\ M_d\ \delta||\mathbf{z}-\mathbf{y}||.$$

Putting everything together we obtain:

$$||\mathbf{f}_\delta(\mathbf{y}) - \mathbf{f}_\delta(\mathbf{z})|| \leq d^{1/2}c_1(1/m_1 + M_d)\ \delta\ ||\mathbf{z}-\mathbf{y}||, \quad C_2 := d^{1/2}c_1(1/m_1 + M_d).$$

Thus if we choose $\delta_0 = \min\{1/2, 1/(2C_1),\ 1/(2C_2)\}$ the proof is over. $\square$

To summarize, we have just constructed some small enough $\delta_0 \leq 1/2$ so that the map

$$\mathbf{f}_{\delta_0}(\mathbf{x}) = \mathbf{x} - [H(\mathbf{x})]^{-1}[\nabla\phi(\mathbf{x})], \quad \forall\mathbf{x}\in\overline{B_\delta(\mathbf{a})}$$

becomes a contraction, thus it has a unique fixed point, which from (3.36) we know that it must equal $\mathbf{a}$.

The advantage is that we can determine/estimate $\mathbf{a}$ by iterating $\mathbf{f}_{\delta_0}$. The only problem is to obtain an initial point which sits a-priori close to the **unknown a**, in this case at a distance less than δ_0.

3.3.1 The main idea of the combined algorithm

1. Let us assume that we have an a-priori information about the m appearing in (3.26). This amounts to obtaining a positive lower bound on the lowest eigenvalue of the Hessian matrix, **everywhere** in space. Since the right hand side of (3.26) only depends on $\mathbf{x}_0$ and m, we can determine the time t_1 we need such that

$$||\mathbf{a} - \mathbf{x}(t_1)|| \leq 1/2.$$

2. Consider the closed ball $T := \overline{B_2(\mathbf{x}(t_1))}$. Clearly,

$$\overline{B_1(\mathbf{a})} \subset T.$$

   Using m instead of m_1 and T instead of $\overline{B_1(\mathbf{a})}$, we can get upper bound estimates for the constants s_1, M_d and c_1 in (3.29), (3.30) and (3.32). From the explicit formulas of C_1 and C_2 given in the previous lemma we see that when we compute them using the upper bounds, we get something larger: $C_1 \leq \tilde{C}_1$ and $C_2 \leq \tilde{C}_2$.

3. Compute the quantity:

$$\delta_1 := \min\{1/2, 1/(2\tilde{C}_1),\ 1/(2\tilde{C}_2)\} \leq \delta_0.$$

4. Run the gradient method again up to a time $t_2 \geq t_1$ such that the right hand side of (3.26) becomes less than δ_1, hence δ_0.

5. Define the sequence:

$$\mathbf{y}_1 := \mathbf{x}(t_2), \quad \mathbf{y}_{n+1} := \mathbf{f}_{\delta_0}(\mathbf{y}_n),\ n \geq 1.$$

From the Banach fixed point theorem we know that this sequence must converge to $\mathbf{a}$, the unique fixed point of $\mathbf{f}_{\delta_0}$. Let us investigate how fast it converges. From (3.36) we have:

$$||\mathbf{y}_{n+1} - \mathbf{a}|| = ||\mathbf{f}_{\delta_0}(\mathbf{y}_n) - \mathbf{a}|| \leq C_1||\mathbf{y}_n - \mathbf{a}||^2.$$

Thus we have:

$$||\mathbf{y}_n - \mathbf{a}|| \leq C_1||\mathbf{y}_{n-1} - \mathbf{a}||^2 \leq C_1^3||\mathbf{y}_{n-2} - \mathbf{a}||^4 \leq ... \leq C_1^{2^{n-1}-1}||\mathbf{y}_1 - \mathbf{a}||^{2^{n-1}},$$

which suggests:

$$||\mathbf{y}_n - \mathbf{a}|| \leq C_1^{-1}(C_1\delta_0)^{2^{n-1}}, \quad n \geq 1, \tag{3.38}$$

inequality which can be proved by induction. From the definition of δ_0 we know that $C_1\delta_0 \leq 1/2$ hence:

$$||\mathbf{y}_n - \mathbf{a}|| \leq C_1^{-1}\frac{1}{2^{2^{n-1}}}, \quad n \geq 1.$$

The convergence in (3.38) is very fast. If, say, $C_1 = 1$ and $\delta_0 = 10^{-1}$, then that after one iteration the error is 10^{-2}, after two iterations is 10^{-4}, and after four iterations is already 10^{-16}.