Stationary Points; Minima and Maxima; Gradient Method

Stationary Points Let $\phi(X)$ be a continuously differentiable function for $X \in \mathcal{D}$, a region in $\mathbf{R}^n$.

Definition A point $X_0 \in \mathcal{D}$ is a *stationary point* for $\phi(X)$ if $\nabla\phi(X_0) = 0$, i.e., if

$$\frac{\partial \phi}{\partial x_k}(X_0) = \frac{\partial \phi}{\partial x_k}(x_{0,1}, x_{0,2}, ..., x_{0,n}) = 0, \;\; k = 1, 2, ..., n.$$

In other words, a stationary point is characterized as being the solution of a certain set of n equations in n unknowns obtained by setting all the partial derivatives of $\phi(X)$ equal to zero.

A stationary point for $\phi(X)$ need not be either a maximum or a minimum for $\phi(X)$; an example is obtained by considering the function

$$\phi(x, y) = x^2 - y^2.$$

We have

$$\frac{\partial \phi}{\partial x}(0,0) = 2x\,|_{x=0} = 2 \cdot 0 = 0; \;\; \frac{\partial \phi}{\partial y}(0,0) = -2y\,|_{y=0} = -2 \cdot 0 = 0$$

but clearly

$$\phi(x, 0) > 0 = \phi(0,0), \;\; x \neq 0; \;\; \phi(0, y) < 0 = \phi(0,0), \;\; y \neq 0.$$

Because of the shape of its graph, such a point is sometimes called a *saddle point.*

Definition A point $X_0 \in \mathcal{D}$ is a (global) minimum for $\phi(X)$ in $\mathcal{D}$ if $\phi(X) \geq \phi(X_0)$ for every $X \in \mathcal{D}$. It is a (global) maximum if the inequality is reversed: $\phi(X) \leq \phi(X_0),\ X \in \mathcal{D}$. If the inequality is only valid in a "neighborhood"

$$\mathcal{N}(X_0, \epsilon) = \{X \in \mathcal{D} \,|\, \|X - X_0\| < \epsilon\},$$

for some $\epsilon > 0$, then we have a *local minimum*, or a local maximum, at X_0.

Clearly a global minimum/maximum is also a local minimum/maximum. So any results we give for the latter also apply to the former.

Theorem A local minimum, or maximum, X_0, for a function $\phi(X)$, continuously differentiable in $\mathcal{D}$, is a stationary point: $\nabla\phi(X_0) = 0$.

Proof It will be enough to treat the case of a local minimum because if $\phi(X)$ has a local maximum at X_0 then $\psi(X) \equiv -\phi(X)$ has a local minimum there and, clearly, $\nabla\psi(X_0) = 0 \Rightarrow \nabla\phi(X_0) = 0$ and vice versa. From our definition of "region", $\mathcal{D}$ is *open*; every point in $\mathcal{D}$ has a neighborhood still lying in $\mathcal{D}$. We find such a neighborhood for X_0:

$$\mathcal{N}(X_0, \rho) = \{X \in \mathcal{D} \,|\, \|X - X_0\| < \rho\}.$$

Then we find $\mathcal{N}(X_0, \epsilon),\ \epsilon \leq \rho$, so that

$$\phi(X_0) \leq \phi(X),\ X \in \mathcal{N}(X_0, \epsilon).$$

We form the line through X_0 in the direction of of the (transposed) gradient, $\nabla\phi(X_0)^*$. It consists of parametrized vectors $X_0 + t\,\nabla\phi(X_0)^*$. For t sufficiently small, these vectors will lie in $\mathcal{N}(X_0, \epsilon)$, so we will have, for such t,

$$\phi(X_0 + t\,\nabla\phi(X_0)^*) \geq \phi(X_0).$$

This can only happen if

$$0 = \frac{d}{dt}\phi(X_0 + t\,\nabla\phi(X_0)^*)\bigg|_{t=0} = \nabla\phi(X_0)\,\frac{d}{dt}\left(X_0 + t\,\nabla\phi(X_0)^*\right)$$

$$= \nabla\phi(X_0)\,\nabla\phi(X_0)^* = \|\nabla\phi(X_0)\|^2,$$

from which we conclude that $\nabla\phi(X_0) = 0$.

This condition allows us to locate possible minima/maxima for the function $\phi(X)$ by solving the $n \times n$ system of equations $\nabla\phi(X) = 0$.

Example 1 Consider the function in $\mathbf{R}^2$:

$$\phi(x, y) = y^4 - 2y^2 + \frac{x^2}{2} + xy + x + y + 1.$$

The gradient vector is

$$\nabla\phi(x, y) = (\, x + y + 1 \quad 4y^3 - 4y + x + 1\,).$$

Setting both components equal to 0 we obtain the system of two equations in two variables

$$x + y + 1 = 0; \;\; 4y^3 - 4y + x + 1 = 0.$$

The first gives $x + 1 = -y$; substituting this in the second we have

$$4y^3 - 5y = 0.$$

Solving this equation for its three solutions and then using $x = -(y+1)$ we obtain the three solution pairs which constitute the stationary points of $\phi(x, y)$:

$$\begin{array}{c} y = 0, \;\; x = -1; \\ y = \frac{\sqrt{5}}{2} = 1.1180, \;\; x = -\frac{\sqrt{5}}{2} - 1 = -2.1180;. \\ y = -\frac{\sqrt{5}}{2} = -1.1180, \;\; x = \frac{\sqrt{5}}{2} - 1 = .1180 \end{array}$$

Then we compute the values

$$\phi(-1, 0) = .5, \quad \phi(-2.1180, 1.1180) = -1.0625,$$

$$\phi(.1180, -1.1180) = -1.0625.$$

Since $|\phi(x, y)|$ clearly goes to ∞ as $\|(x, y)\| \to \infty$, the function $\phi(x, y)$ must have at least one minimum. We conclude from the gradient analysis that there must, in fact, be two; one at $(-2.1180, 1.1180)$, the other at $(.1180, -1.1180)$. In both cases the value is -1.06250.

What can we say about the point $(0, -1)$? Is it a (necessarily local) minimum, a maximum or just a stationary point? We compute the second order partial derivatives

$$\frac{\partial^2 \phi}{\partial x^2}(-1,0) \;=\; 1; \;\; \frac{\partial^2 \phi}{\partial y^2}(-1,0) \;=\; \big(12\,y^2 - 4\big)\big|_{x=-1,\,y=0} \;=\; -4.$$

From this it is clear that ϕ increases as x moves slightly away from the value -1, keeping y fixed at 0, while ϕ decreases as y moves slightly away from the value 0, keeping x fixed at -1. We conclude $(-1, 0)$ is neither a maximum nor a minimum; it is just a stationary point. In our section on the **Hessian matrix** of second order partial derivatives of a scalar valued function $\phi(X)$ we will extend these ideas to a more comprehensive analysis.

Many "real world" applications involve finding minima or maxima of a function $\phi(X)$; for example in cases where $\phi(X)$ represents the efficiency of a process, the cost of an operation, etc. Consequently, it is important to have systematic procedures for finding points where minima or maxima are achieved. We have seen that one way to do this is to solve the system of equations, $\nabla\phi(X) = 0$ obtained by setting the gradient equal to zero, i.e.,

$$\begin{aligned} \tfrac{\partial \phi}{\partial x_1}(x_1, x_2, ..., x_n) \;&=\; 0, \\ \tfrac{\partial \phi}{\partial x_2}(x_1, x_2, ..., x_n) \;&=\; 0, \\ &\cdots \\ \tfrac{\partial \phi}{\partial x_n}(x_1, x_2, ..., x_n) \;&=\; 0. \end{aligned}$$

In most cases this system cannot be solved explicitly; one needs to resort to equation solving techniques such as Newton's method, about which we have more to say elsewhere.

Another method commonly used for finding minima, at least in principle, involves following a *path of steepest descent*; in the case of a maximum this would be replaced by a *path of steepest ascent*. We recall that, for a unit vector U, the *directional derivative* of $\phi(X)$, at X_0, in the direction U, is defined as

$$\frac{\partial \phi}{\partial U}(X_0) \;=\; \nabla\phi(X_0)\,U.$$

The **Schwartz inequality** implies, since $\|U\| \;=\; 1$,

$$\left|\frac{\partial \phi}{\partial U}(X_0)\right| \;=\; |\nabla\phi(X_0)\,U| \;\le\; \|\nabla\phi(X_0)\|\;\|U\| \;=\; \|\nabla\phi(X_0)\|\,.$$

If we define $\hat{U} \;=\; \frac{1}{\|\nabla\phi(X_0)\|}\,\nabla\phi(X_0)^*$, the normalized column vector version of the gradient of ϕ at X_0, which we term the *gradient direction*, we see that

$$\frac{\partial \phi}{\partial \hat{U}}(X_0) \;=\; \frac{\nabla\phi(X_0)\,\nabla\phi(X_0)^*}{\|\nabla\phi(X_0)\|} \;=\; \|\nabla\phi(X_0)\|\,;$$

the direction $\hat{U}$ thus maximizes the directional derivative. In the same way we can see that the negative gradient direction $-\hat{U}$ minimizes the directional derivative, with

$$\frac{\partial \phi}{\partial(-\hat{U})} \;=\; -\,\frac{\partial \phi}{\partial \hat{U}}(X_0) \;=\; -\,\|\nabla\phi(X_0)\|\,.$$

For the *gradient field* $F(X) \;=\; \nabla\phi(X)^*$ the associated *streamlines* are the solutions of the vector system of differential equations

$$\frac{dX}{dt} \;=\; \nabla\phi(X(t))^*, \text{ i.e.,}$$

$$\frac{dx_k}{dt} \;=\; \frac{\partial \phi}{\partial x_k}\,(x_1(t), x_2(t), ..., x_n(t))\,,\;\; k \;=\; 1, 2, ..., n.$$

The unit tangent vector to a solution $X(t)$ of this system is $\hat{U} \;=\; \frac{1}{\|\nabla\phi(X(t))\|}\nabla\phi(X(t))$ and thus points in the direction of steepest ascent for $\phi(X)$. We can compute

$$\frac{d}{dt}\phi(X(t)) \;=\; \nabla\phi(X(t))\nabla\phi(X(t))^* \;=\; \|\nabla\phi(X(t))\|^2\,.$$

Following $X(t)$ for increasing t results in increasing $\phi(X(t))$; following $X(t)$ for decreasing values of t results in decreasing $\phi(X(t))$. As $t \to \infty$ streamlines $X(t)$ either go off to infinity in $\mathbf{R}^n$, or tend to a maximum of $\phi(X)$, or approach a saddle point of $\phi(X)$ in an increasing direction (from a probabilistic standpoint the latter is very unlikely but should be included for completeness).

In the same way, as $t \to -\infty$ streamlines $X(t)$ either go off to infinity, approach a minimum of $\phi(X)$ or approach a saddle point from a decreasing direction (the latter again unlikely). In practice, to find a minimum this way, we reverse the "time sense" t and consider the system

$$\frac{dX}{dt} = -\nabla\phi(X(t))^*,$$

along whose solutions, $X(t)$, $\phi(X(t))$ decreases. Thus following streamlines presents one possible method for locating maxima and/or minima. In some cases the solutions $X(t)$ can be computed explicitly and the maxima/minima located by finding points $\hat{X}$ which have the form $\hat{X} = \lim_{t\to\pm\infty} X(t)$.

However, this is rarely the case, particularly in significant applications, and one usually needs to resort to numerical procedures. A common numerical procedure for approximating solutions $X(t)$ of a system

$$\frac{dX}{dt} = F(X(t))$$

is known as *Euler's Method*; if we discretize the t variable via points t_k, $-\infty < k < \infty$, with $t_{k+1} = t_k + h$, for some fixed *step length* h, and then approximate the derivative by a difference quotient, we have

$$\frac{1}{h}\,(X(t_{k+1}) - X(t_k)) \approx F(X(t_k)).$$

It makes sense, therefore, to generate approximations X_k to $X(t_k)$ via

$$\frac{1}{h}\,(X_{k+1} - X_k) = F(X_k),\ \ k = 0, 1, 2, ...,$$

starting at some chosen point X_0, preferably close to the point $\hat{X}$ we are searching for. Taking $F(X) = \nabla\phi(X)^*$, which is our real interest here, and solving for the point X_{k+1}, we have

$$X_{k+1} = X_k + h\,\nabla\phi(X_k)^*,\ \ k = 0, 1, 2, ...$$

With X_0 skillfully chosen and h sufficiently small, we can expect $\lim_{k\to\infty} X_k = \hat{X}$, ordinarily a maximum in this setting since we are generating approximations to $X(t)$ for increasing values of t.

On the other hand, solutions of

$$\frac{dX}{dt} = -\nabla\phi(X(t))^*$$

have the same solutions $X(t)$ as the earlier system, but oriented in the opposite direction with respect to t; in this case $\phi(X(t))$ decreases as $t \to \infty$. Applying Euler's Method here we obtain

$$X_{k+1} = X_k - h\,\nabla\phi(X_k)^*,\ \ k = 0, 1, 2, ...$$

This recursive equation is what we call the *gradient method* for finding minima of $\phi(X)$; the earlier recursive system $X_{k+1} = X_k + h\,\nabla\phi(X_k)^*$, is the gradient method for finding maxima.

Whether looking for maxima or for minima, in using the gradient method the idea is to choose a point X_0 which one suspects to be close to the maximum/minimum one is looking for. The gradient method equations are then used to generate a sequence of points $X_k,\ k = 1, 2, 3, ...$ in the expectation that they will converge, as $k \to \infty$ to the desired point $\hat{X}$. While one typically works with a modestly small value of h, say $h = .1$, e.g., it may, in practice be necessary to take h quite small to ensure convergence.

Example 2 Let us consider the function

$$\phi(x,\, y) \;=\; x^2 \,+\, 2\,x\,y \,+\, 3\,y^2 \,-\, 2\,x \,+\, 3\,y,$$

for which the gradient, as a function of x and y, is

$$\nabla\phi(x,\, y) \;=\; (\,2\,x \,+\, 2\,y \,-\, 2\,,\quad 2\,x \,+\, 6\,y \,+\, 3\,)\,.$$

Setting the gradient equal to zero we have the two equations

$$2\,x \,+\, 2\,y \,-\, 2 \,=\, 0, \qquad 2\,x \,+\, 6\,y \,+\, 3 \,=\, 0,$$

for which the solution, which is the global minimum, is easily seen to be $\hat{x} \;=\; 9/4,\;\; \hat{y} \;=\; -5/4$. In this case the gradient method for minimization reads

$$\begin{pmatrix} x_{k+1} \\ y_{k+1} \end{pmatrix} \;=\; \begin{pmatrix} x_k \\ y_k \end{pmatrix} \;-\; h\, \begin{pmatrix} 2\,x_k \,+\, 2\,y_k \,-\, 2 \\ 2\,x_k \,+\, 6\,y_k \,+\, 3 \end{pmatrix}.$$

Taking $x_0 \;=\; .5,\;\; y_0 \;=\; -1,\;\; h \;=\; .2$ and applying this method for 10 steps yields the table

$$\begin{pmatrix} k & x_k & y_k \\ 1 & 1.1 & -.6 \\ 2 & 1.3 & -.92 \\ 3 & 1.548 & -.936 \\ 4 & 1.7032 & -1.032 \\ 5 & 1.8347 & -1.0749 \\ 6 & 1.9308 & -1.1189 \\ 7 & 2.0060 & -1.1485 \\ 8 & 2.0630 & -1.1727 \\ 9 & 2.1069 & -1.1907 \\ 10 & 2.1404 & -1.2046 \end{pmatrix}.$$

In general the method is slow but computationally useful when the equations obtained by setting the gradient equal to 0 are hard or impossible to solve. Even when these equations can be solved this *iterative* method provides a valuable check on the result.