The Big Valley

Jarrod Barkley is questioning his younger brother Nick on the mule Nick just bought. “Nick when this mule is loaded down with all the gear we need to take, it can only walk downhill and only in straight lines. Every time we need to turn, we are going to have to unload the mule, turn him to new the new direction and load him up again.”

The time is the 1870, the location is California’s San Joaquin Valley and the business is mining. It is one of the many businesses own by Victoria Barkley. Victoria is the mother of both Jarrod and Nick, and she also has a daughter Audra. She also accepts as a son the man Health, who the illegitimate son of Victoria’s late husband Tom.

Nick Barkley, the hot-headed younger son, defends his purchase. “That is the beauty of the mule. His name is ‘Gradient Descent’ by the way. Why Jarrod, I would expect you to remember more of you multi-variable calculus. All we need to do is point the mule in the negative direction of the gradient and let him go. He will stop right at the bottom. I even checked that it works for the equation $f(x, y) = x^2 + y^2$. No matter where we start in this valley, the mule will go directly to lowest point at $(0, 0)$ and stop. There will not be any reloading of the mule”

Jarrod replies, “Nick your function $f(x, y)$ is much too special. The contour lines of the valley are more likely to be ovals than circles. And for an elliptic contours, say for a function like $f(x, y) = x^2/a^2 + y^2/b^2$ the negative gradient line will not pass through the origin except for the special points along the axes. We will have to turn the mule, perhaps a number of times, has we go towards the bottom.”

Before continuing our story lets develop some notation. We will be looking at functions $f(x, y)$ of two variables and we will be searching for (global) minimum value of f. Our search process will be iterative, that is, it will take an approximation a point $P = P_i = (x_i, y_i)$, and then compute the “next approximation point” $Q = P_{i+1} = (x_{i+1}, y_{i+1})$. In the most general case, we will be given a direction $\vec{u} = \langle u_1, u_2\rangle$ in addition to the starting point P.

Given $P = (p, q)$ and $\vec{u} = \langle a, b\rangle$ the *$P\vec{u}$-line* is the line with parametric equations

$$x = p + at \qquad y = q + bt$$

and the *$P\vec{u}$-section* is the function of one variable t given by

$$g(t) = f(p + at, q + bt)$$

which is the restriction of f over the $P\vec{u}$-line. The *next $P\vec{u}$-approximation*, is the local minimum of g adjacent to $t = 0$, in the direction that g decreases. (So if $g'(0) = f_{\vec{u}}(P) < 0$, this be in the direction of $t > 0$, if $g'(0) > 0$, this will be the direction $t < 0$ and if $g'(0) = 0$ it could be P itself. Often, P will be understood and we will just call Q the *next approximation in the direction of $\vec{u}$.* Often $\vec{u}$ will be understood to be $-\nabla f(P)$, the negative of the gradient of f at P, we will just call Q the *next approximation.* So like the mule, we only go “downhill” and we stop as soon as it starts to go “uphill” again.

This method, when $\vec{u} = -\nabla f(P)$, is called *Gradient Descent* or *Steepest Descent.* It has the advantage, that each “next approximation” is in the direction of the greatest decrease of f. But it has disadvantages too, as we will shortly see. One of the disadvantages of this method is that it can get trapped by a local minimum M which is not the global minimum. If the current point P is near enough to M all the next approximations will be near M. Each local minimum has a region of attraction, a “basin” of points that “drain” to P. But that is not part of this project.

1. *Show that both Nick’s and Jarrod’s statements are both correct.* Let $f(x, y) = Ax^2 + By^2$ where $A > 0$ and $B > 0$. Let $P = (p, q)$ with not both p and q zero, and let $\vec{u} = -\nabla f(P)$. Show that the $P\vec{u}$ line goes through the origin, the minimum of f, when $A = B$ or when $A \neq B$ and one of p and q is zero. In this case show that the next approximation is the origin. Show if $A \neq B$ and both $p > 0$ and $q > 0$, then the $P\vec{u}$-line misses the origin completely, it does not pass through the origin, so the next approximation is not the origin.
2. *A Theory Question* Suppose P is given $\vec{u} = -\nabla f(P)$ and the next approximation is Q. Show that $P\vec{u}$-line and the level curve $f(x, y) = f(Q)$ are tangent at Q. Conclude that $\vec{u}$ and $-\nabla f(Q)$ are perpendicular.

> Back to our story, Audra, who has been looking over her brother's shoulders speaks up. "Jarrod has a point, the path will be a zig-zag route and it looks like it could zig-zag more than a little. Why don't you just find the local minimum and just point the mule in that direction to start with?" Nick and Jarrod explain that the valley in question is both subject to the infamous thick San Joaquin tule fogs and that the Rosenbrock valley (see #6 below) curves around huge mountains so that the global minimum may not be visible even on a clear day.

3. *Zig-zags* Illustrate the zig-zag path for the function $f(x,y)=x^2+2y^2$ starting at location $P=P_0(x_0,y_0)=(1,1)$. Exactly (fractions not decimals) compute the next 5 approximations. Plot your zig-zag together with the contours of f that pass through these approximations. Be sure to have line segments connecting the approximations in order constructed.
4. *A more general, but still quadratic f* Let $f(x,y)=(2x-y)^2+(1+y)^2$. Show f has one local minimum which must be a global minimum (why?). Start at $P=P_0(0,0)$ and compute the next 5 approximations exactly (fractions not decimals). Plot your zig-zags and contours like in #3.
5. *The Rosenbrock Valley* The Rosenbrock valley has as a function the Rosenbrock function, $f(x,y)=100(y-x^2)^2+(1-x)^2$ which is surprising hard to draw. Find and classify all the critical points of f. Find the coordinates (a,b) where f has a global minimum. Find a way to plot f over the range x=a-1.5..a+1.5 and y=b-1.5..b+1.5 that shows that it is a steep valley but with a gentle slope inside the valley. Here plot3d shows a surface that is mostly gentle, while a more narrow view given by the surface

$$\vec{r}(s,t)=\left\langle s, s^2+t, f(s,s^2+t)\right\rangle \qquad s=a-1.5..a+1.5, \quad t=-L..L$$

shows the steep sides, but not the gentle slope inside the valley. A contour plot with a very careful choice of contours is needed.
6. *More Rosenbrock Valley* For a fixed $x=t$, show that $f(t,y)$ has a single local minimum when $y=t^2$, but the gradient of f at this point, namely $\nabla f(t,t^2)$ points parallel to the x-axis and not in the direction of the tangent of the curve $\vec{r}(t)=\left\langle t,t^2\right\rangle$, (when $t\neq 0$). Using $P=P_0(-1,1)$ as a starting point for gradient descent show the next four approximations are located. Use decimals and not fractions.

> Meanwhile back at the ranch, The thoughtful brother, Health notices "Jarrod it looks like your oval case has a better solution. Instead of going in the direction of the gradient, we could first go in a direction parallel to one of the axes of the ellipses, then the right angle zig-zag turn would be parallel to the second axes and the second step would end up at the local minimum. Of course we would have to know the directions of the axes of the ellipses before hand"
>
> Victoria happens to be walking by at this moment. Victoria reminds one of Barbara Stanwyck, in both mind and body. She reflects "Actually boys, given any direction, $\vec{u}$, not just one parallel to the axis, there is an easy to compute "conjugate direction" $\vec{v}$ for the second step which will always take you directly to the local minimum of your oval case. So you could use the gradient direction for the first step, and take the conjugate direction for the second. Of course, if the function is not a quadratic, in may not arrive at the local minimum in two steps.

Lets update the notation. The *conjugate direction* to a vector $\vec{u}=\langle u_1,u_2\rangle$ is any non-zero vector $\vec{v}=\langle v_1,v_2\rangle$ so that the expression $au_1v_1+bu_1v_2+bu_2v_1+cu_2v_2=0$. The scalars a,b and c are the values of the second partials $a=f_{xx}(P),b=f_{xy}(P)$ and $c=f_{yy}(P)$ and so depend both on the function f and the point P. An approximation that alternates between the negative gradient $\vec{u}$ and the conjugate to $\vec{u}$ is an an example of a *Conjugate Gradient Method.*
7. *Show Health is correct* For problem #3 starting with the same $P(1,1)$ compute the next approximation in the $-\hat{\imath}$ direction, and then the next approximation in the $-\hat{\jmath}$-direction. Show the same for problem #4 after you figure out the axes of those ellipses.
8. *Conjugate Gradient* Go back to problems #3 and #4. Computer the conjugate direction $\vec{v}$ to first $\vec{u}$ and show that the second approximation, using $\vec{v}$ instead of $-\nabla f(P_1)$, jumps directly to the global minimum.

From the course Syllabus

BIG PROJECT. You will work on the project in groups of 1–4 students. This project will be a substantial assignment, giving you a chance to earn part of your grade in an environment which simulates the so-called "real world" better than does an in-class exam. It will also give your instructor a chance to base part of your grade on your best work, produced in a setting where time should not be a factor (assuming you start on your project as soon as it is assigned). The results of your work on your project will be presented in a report (one report per group). Each member will also submit a "group evaluation" giving their impression of the relative contribution of each member to the group's effort. These evaluations are due with the project. It is not guaranteed that each member of the group will receive the same grade. The reports will be graded not only on their mathematical content but also on the QUALITY of the presentation: CLARITY, NEATNESS, and PROPER GRAMMAR are also important. Both reports and group evaluations must be **TYPED**. The project was be assigned on Thursday, 23 Oct and due on Thursday, 6 Nov

GRADING. The big project is worth 100 points or five times the value of an ordinary project but only twice the work time. Mathematical content is worth 80% which leaves 20 points for clarity, neatness, grammar and general wow value, for a total of 100 points. Pride of ownership is one of the keys in the subjective points department. You are free to use Maple on all the calculations, but for clarity you should do the easy ones by hand. There are a very small number of bonus points.

The web page for the project, which has TI-89 and Maple code and an worked example with graph, is located at this URL:

http://www.math.fsu.edu/∼bellenot/class/f03/cal3/project.html