Gradient descent method

1. Introduction

The gradient of f at x_0, denoted $\nabla f(x_0)$, if it is not a zero vector, is orthogonal to the tangent vector to an arbitrary smooth curve passing through x_0 on the level set $f(x)=c$. Showed as the picture below:

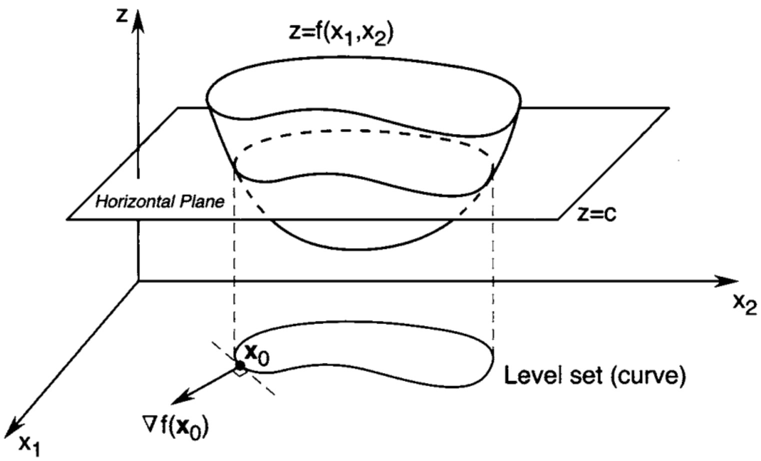

Figure 1 Constructing a level set corresponding to level c for f

Thus, the direction of maximum rate of increase of a real-valued differentiable function at a point is orthogonal to the level set of the function through that point. In other words, the gradient acts in such a direction that for a given small displacement, the function f increases more in the direction of the gradient than in any other direction.

Proof:

Recall that $\langle \nabla f(x), d\rangle, \|d\|=1$, is the rate of increase of f in the direction d at the point x. By the Cauchy-Schwarz inequality,

$$\langle \nabla f(x), d\rangle \le \|\nabla f(x)\|$$

Because $\|d\|=1$. But if $d=\nabla f(x)/\|\nabla f(x)\|$, then

$$\left\langle \nabla f(x), \frac{\nabla f(x)}{\|\nabla f(x)\|}\right\rangle = \|\nabla f(x)\|$$

Thus, the direction in which $\nabla f(x)$ points is the direction of maximum rate of increase of f at x. The direction in which $-\nabla f(x)$ points is the direction of maximum rate of decrease of f at x. Hence, the direction of negative gradient is a good direction to search if we want to find a function minimizer.

Let $x^{(0)}$ be a starting point, and consider the point $x^{(0)}-\alpha\nabla f(x^{(0)})$. Then, by Taylor's theorem, we obtain

$$f(x^{(0)}-\alpha\nabla f(x^{(0)})) = f(x^{(0)}) - \alpha\left\|\nabla f(x^{(0)})\right\|^2 + o(\alpha)$$

Thus, if $\nabla f(x^{(0)})\neq 0$, then for sufficiently small $\alpha>0$, we have

$$f(x^{(0)}-\alpha\nabla f(x^{(0)})) < f(x^{(0)})$$

This means the point $x^{(0)} - \alpha \nabla f(x^{(0)})$ is an improvement over the point $x^{(0)}$ if we are searching for a minimizer.

To formulate an algorithm that implements this idea, suppose that we are given a point $x^{(k)}$ .To find the next point $x^{(k+1)}$, we start at $x^{(k)}$ and move by an amount $-\alpha_k \nabla f(x^{(0)})$ where α_k is a positive scalar called the *step size*. This procedure leads to the following iterative algorithm:

$$x^{(k+1)} = x^{(k)} - \alpha_k \nabla f(x^{(k)})$$

We refer to this as a gradient descent algorithm (or simply a gradient algorithm). The gradient varies as the search proceeds, tending to zero as we approach the minimizer.