(limingkuan2017@outlook.com)

Gradient descent method

CONTENTS

1. Introduction 1
2. The Method of Steepest Descent 2

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:

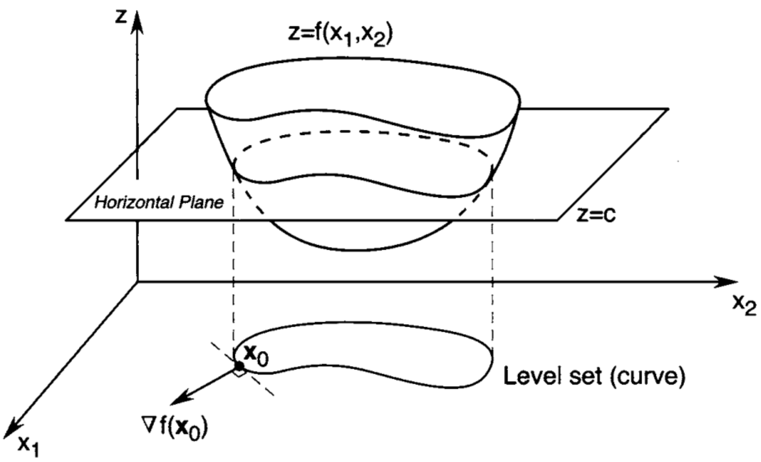

Figure 1 Constructing a level set corresponding to level c for f

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

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

(limingkuan2017@outlook.com)
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.

2. The Method of Steepest Descent

The method of steepest descent is a gradient algorithm where the step size α_k is chosen to achieve the maximum amount of decrease of the objective function at each individual step. Specifically, α_k is chosen to minimize $\phi_k(\alpha) \triangleq f(x^{(k)} - \alpha \nabla f(x^{(k)}))$. In the other words,

$$\alpha_k = \arg\min_{\alpha \geq 0} f(x^{(k)} - \alpha \nabla f(x^{(k)}))$$

To summarize, the steepest descent algorithm proceeds as follows: At each step, starting from the point $x^{(k)}$, we conduct a line search in the direction $-\nabla f(x^{(k)})$ until a minimizer, $x^{(k+1)}$ is found. A typical resulting from the method of steepest descent is depicted in the figure below:

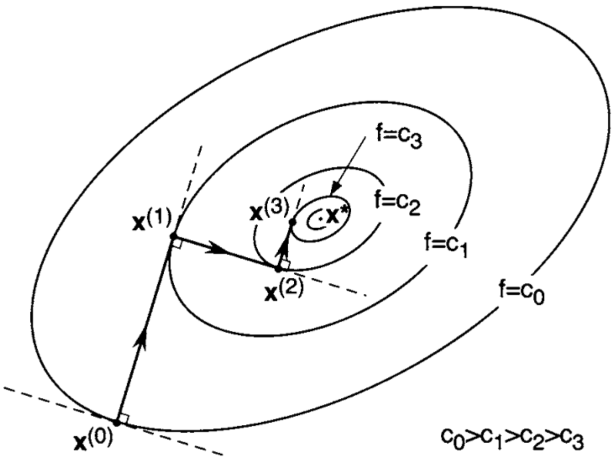

Figure 2 Typical sequence resulting from the method of steepest descent.

Observe that the method of steepest descent moves in orthogonal steps, as stated in the following proposition.

Proposition: if $\{x^{(k)}\}_{k=0}^{\infty}$ is a steepest descent sequence for a given function $f: R^n \to R$, then for each k the vector $x^{(k+1)} - x^{(k)}$ is orthogonal to the vector $x^{(k+2)} - x^{(k+1)}$.

Proof. From the iterative formula of the method of steepest descent it follows that:

$$\left\langle x^{(k+1)} - x^{(k)}, x^{(k+2)} - x^{(k+1)} \right\rangle = \alpha_k \alpha_{k+1} \left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle$$

To complete the proof, it is enough to show that

$$\left\langle \nabla f(x^{(k)}), \nabla f(x^{(k+1)}) \right\rangle = 0$$

Observe that α_k is a nonnegative scalar that minimizes $\phi_k(\alpha) \triangleq f(x^{(k)} - \alpha \nabla f(x^{(k)}))$. Hence, using the FONC and the chain rule gives us

$$\begin{aligned} 0 &= \phi_k'(\alpha_k) \\ &= \frac{d\phi_k}{d\alpha}(\alpha_k) \\ &= \nabla f(x^{(k)} - \alpha_k \nabla f(x^{(k)}))^T (-\nabla f(x^{(k)})) \\ &= -\left\langle \nabla f(x^{(k+1)}), \nabla f(x^{(k)}) \right\rangle \end{aligned}$$

Which completes the proof.

And we can proof that if $\{x^{(k)}\}_{k=0}^{\infty}$ is the steepest descent sequence for $f: R^n \to R$ and if $\nabla f(x^{(k)}) \neq 0$, then $f(x^{(k+1)}) < f(x^{(k)})$. If for some k, we have $\nabla f(x^{(k)}) = 0$, then the point $x^{(k)}$ satisfies the FONC. In this case, $x^{(k+1)} = x^{(k)}$. We can use the above as the basis for a stopping(termination) criterion for the algorithm.

The condition $\nabla f(x^{(k+1)}) = 0$, however is not directly suitable as a practical stopping criterion, because the numerical computation of the gradient will rarely be identically equal to zero. Alternatively, we may compute the absolute difference $\left| f(x^{(k+1)}) - f(x^{(k)}) \right|$ between objective function values for every two successive iterations, and if the difference is less than some prespecified threshold, then we stop; that is, we stop when

$$\left| f(x^{(k+1)}) - f(x^{(k)}) \right| < \varepsilon$$

Yet another alternative is to compute the norm $\left\| x^{(k+1)} - x^{(k)} \right\|$ of the difference between two successive iterates, and we stop if the norm is less than a prespecified threshold:

$$\left\| x^{(k+1)} - x^{(k)} \right\| < \varepsilon$$

Alternatively, we may check "relative" values of the quantities above; for example,

$$\frac{\left| f(x^{(k+1)}) - f(x^{(k)}) \right|}{\left| f(x^{(k)}) \right|} < \varepsilon$$

Or

$$\frac{\left\| x^{(k+1)} - x^{(k)} \right\|}{\left\| x^{(k)} \right\|} < \varepsilon$$

The two relative stopping criteria above are preferable to the absolute criteria because the relative criteria are "scale-independent." To avoid dividing by very small numbers, we can modify these stopping criteria as follows:

$$\frac{\left| f(x^{(k+1)}) - f(x^{(k)}) \right|}{\max\{1, \left| f(x^{(k)}) \right|\}} < \varepsilon$$

Or

$$\frac{\left\| x^{(k+1)} - x^{(k)} \right\|}{\max\{1, \left\| x^{(k)} \right\|\}} < \varepsilon$$

Example, we use the method of steepest descent to find the minimizer of:

(limingkuan2017@outlook.com)

$$f(x_1,x_2,x_3)=(x_1-4)^2+(x_2-3)^2+(x_3+5)^2$$

The initial point is $x^{(0)}=[4,2,-1]^T$. We perform three iterations.

First, the gradient of $f(x_1,x_2,x_3)$ is that

$$\nabla f(x)=[4(x_1-4)^3,2(x_2-3),16(x_3+5)^3]^T$$

Hence,

$$\nabla f(x^{(0)})=[0,-2,1024]^T$$

To compute $x^{(1)}$, we need

$$\begin{aligned}\alpha_0&=\arg\min_{\alpha\ge0} f(x^{(0)}-\alpha\nabla f(x^{(0)}))\\&=\arg\min_{\alpha\ge0}(0+(2+2\alpha-3)^3+4(-1-1024\alpha+5)^4)\\&=\arg\min_{\alpha\ge0}\phi_0(\alpha)\end{aligned}$$

Using the **secant method**, we obtain

$$\alpha_0=3.967\times10^{-3}$$

We can draw the function $\phi_0(\alpha)$:

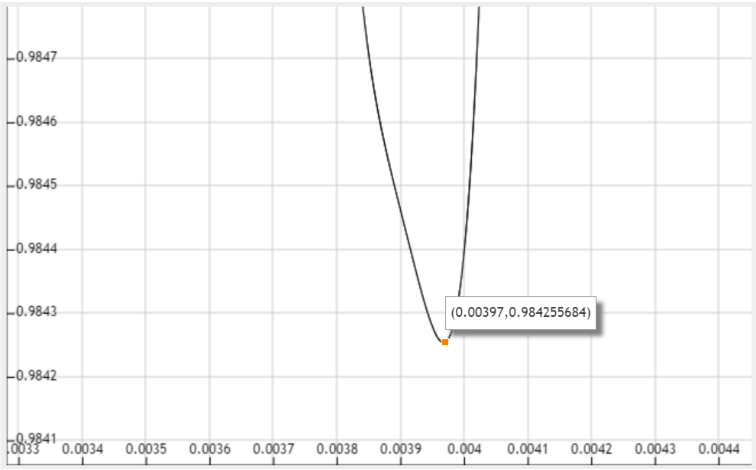

Figure 3 function picture

Secant method please check another report.

Thus,

$$x^{(1)}=x^{(0)}-\alpha_0 f(x^{(0)})=[4.000,2.008,-5,062]^T$$

To find $x^{(2)}$, we first determine

$$\nabla f(x^{(1)})=[0.000,-1.984,-0.003875]^T$$

Next, we find α_1, where

$$\begin{aligned}\alpha_1&=\arg\min_{\alpha\ge0}(0+(2.008+1.984\alpha-3)^2+4(-5.062+0.003875\alpha+5)^4)\\&=\arg\min_{\alpha\ge0}\phi_1(\alpha)\end{aligned}$$

Using the secant method again, we obtain α_1 = 0.5000. Thus,

$$x^{(2)}=x^{(1)}-\alpha_1\nabla f(x^{(1)})=[4.000,3.000,-5.060]^T$$

To find $x^{(3)}$, we can obtain $\alpha_2=16.29$. The value of $x^{(3)}$ is

$$x^{(3)}=[4.000,3.000,-5.002]^T$$

Note that the minimizer of f is $[4,3,-5]^T$, and hence it appears that we have arrived at the minimizer in only three iterations.

Let us now see what the method of steepest descent does with a quadratic function of the form

$$f(x)=\frac{1}{2}x^TQx-b^Tx$$

Where $Q\in R^{n\times n}$ is a symmetric positive definite matrix, $b\in R^n$ and $x\in R^n$. The unique minimizer of f can be found by setting the gradient of f to zero, where

$$\nabla f(x)=Qx-b$$

Because $D(x^TQx)=x^T(Q+Q^T)=2x^TQ$, and $D(b^Tx)=b^T$. There is no loss of generality in assuming Q to be a symmetric matrix. For if we are given a quadratic form $x^TAx(A\neq A^T)$, then because the transposition of a scalar equals itself, we obtain

$$(x^TAx)^T=x^TA^Tx=x^TAx$$

Hence,

$$\begin{aligned} x^TAx &= \frac{1}{2}x^TAx+\frac{1}{2}x^TA^Tx \\ &= \frac{1}{2}x^T(A+A^T)x \\ &\triangleq \frac{1}{2}x^TQx \end{aligned}$$

Note that,

$$(A+A^T)^T=Q^T=A+A^T=Q$$

The Hessian of f is $F(x)=Q=Q^T>0$. To simplify the notation, we write $g^{(k)}=\nabla f(x^{(k)})$. Then, the steepest descent algorithm for the quadratic function can be represented as

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

Where,

$$\begin{aligned} \alpha_k &= \underset{\alpha\geq 0}{\arg\min}\ f(x^{(k)}-\alpha g^{(k)}) \\ &= \underset{\alpha\geq 0}{\arg\min}\left(\frac{1}{2}(x^{(k)}-\alpha g^{(k)})^TQ(x^{(k)}-\alpha g^{(k)})-(x^{(k)}-\alpha g^{(k)})^Tb\right) \end{aligned}$$

Let α_k is a minimizer of $\phi_k(\alpha)=f(x^{(k)}-\alpha g^{(k)})$, we apply the FONC to $\phi_k(\alpha)$ to obtain,

$$\phi_k'(\alpha)=(x^{(k)}-\alpha g^{(k)})Q(-g^{(k)})-b^T(-g^{(k)})=0$$

Hence,

$$\alpha_k=\frac{g^{(k)T}g^{(k)}}{g^{(k)T}Qg^{(k)}}$$

In summary, the method of steepest descent for the quadratic takes the form

$$\boxed{x^{(k+1)}=x^{(k)}-\frac{g^{(k)T}g^{(k)}}{g^{(k)T}Qg^{(k)}}g^{(k)}}$$

Where

$$g^{(k)}=\nabla f(x^{(k)})=Qx^{(k)}-b$$

Example, let

$$f(x_1, x_2) = x_1^2 + x_2^2$$

Then, starting from an arbitrary initial point $x^{(0)} \in R^2$, we arrive at the solution $x^* = 0 \in R^2$ in only one step. See figure below:

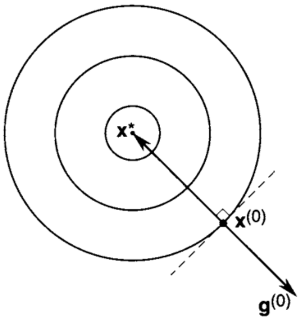

Figure 4 example 1

However, if

$$f(x_1, x_2) = \frac{x_1^2}{5} + x_2^2$$

then the method of steepest descent shuffles ineffectively back and forth when searching for the minimizer in a narrow valley (see Figure 5).

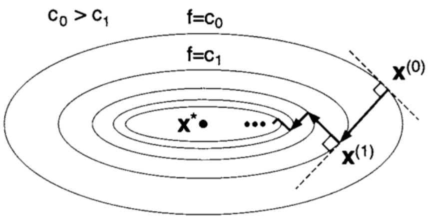

Figure 5 example 2

This example illustrates a major drawback in the steepest descent method. More sophisticated methods that alleviate this problem are discussed in other methods.