Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 3, Lecture 3: Method of Steepest Descent

April 19, 2019 *University of Illinois at Urbana-Champaign*

1 The method of steepest descent

Today we are working with a slightly different set of assumptions. We're going to assume that minimizing a single-variable function is easy (after all, you just have to decide to go left or go right on the number line). Minimizing functions of many variables is hard. So we want a method to reduce the many-variable case to the single-variable case.

The method of steepest descent does precisely this. At each iterative step, we

- Pick a direction to go in
- Solve a one-variable problem to determine how far to go in that direction.

For all of today's class, we will assume that $f : \mathbb{R}^n \to \mathbb{R}$ is a function with a continuous gradient, so that we can take directional derivatives of it.

Recall that under this hypothesis, if $\mathbf{u} \in \mathbb{R}^n$ is a unit vector, then the rate of change of f at $\mathbf{x}$ in the direction of $\mathbf{u}$ is the directional derivative given by

$$\nabla f(\mathbf{x}) \cdot \mathbf{u}.$$

By the Cauchy–Schwarz inequality, we have an upper bound on the size of this directional derivative:

$$|\nabla f(\mathbf{x}) \cdot \mathbf{u}| \le \|\nabla f(\mathbf{x})\| \cdot \|\mathbf{u}\|.$$

This is achieved when $\mathbf{u}$ is parallel to $\nabla f(\mathbf{x})$. To make the directional derivative as negative as possible, we set $\mathbf{u} = -\frac{\nabla f(\mathbf{x})}{\|\nabla f(\mathbf{x})\|}$.

Now we can give a more precise description of the iterative step of the method of steepest descent. Given a point $\mathbf{x}^{(k)} \in \mathbb{R}^n$, we compute the next point $\mathbf{x}^{(k+1)}$ as follows:

1. Compute $\nabla f(\mathbf{x}^{(k)})$.
2. Set $\phi_k(t) = f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)}))$. That is, ϕ_k evaluates f along the line through $\mathbf{x}^{(k)}$ in the direction of steepest descent.
3. Let t_k be the global minimizer of $\phi_k(t)$. This t_k tells us how far along the line we want to go.
4. Go that far along the line: set

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)}).$$

[1] This document comes from the Math 484 course webpage: https://faculty.math.illinois.edu/~mlavrov/courses/484-spring-2019.html

1.1 A quadratic example

Suppose we are trying to minimize $f(x,y) = 4x^2 - 4xy + 2y^2$. The gradient is

$$\nabla f(x,y) = \begin{bmatrix} 8x - 4y \\ -4x + 4y \end{bmatrix}.$$

Because we know how to minimize quadratic functions, we actually know that the minimum is going to be at the only critical point, $(0,0)$. But let's start from the initial guess $\mathbf{x}^{(0)} = (2,3)$ and see how we do.

At $\mathbf{x}^{(0)}$, the gradient is $(4,4)$, so we set

$$\phi_0(t) = f(2-4t, 3-4t) = 4(2-4t)^2 - (2-4t)(3-4t) + 2(3-4t)^2.$$

To minimize $\phi_0(t)$, we compute its derivative:

$$\phi_0'(t) = \nabla f(2-4t, 3-4t) \cdot \nabla f(\mathbf{x}^{(0)}) = \begin{bmatrix} 8(2-4t) - 4(3-4t) \\ -4(2-4t) + 4(3-4t) \end{bmatrix} \cdot \begin{bmatrix} 4 \\ 4 \end{bmatrix} = 64t - 32.$$

Since $\phi_0'(t) < 0$ when $t < \frac{1}{2}$ and $\phi_0'(t) > 0$ when $t > \frac{1}{2}$, the global minimizer is at $t_0 = \frac{1}{2}$.

We set $\mathbf{x}^{(1)} = \mathbf{x}^{(0)} - \frac{1}{2}\nabla f(\mathbf{x}^{(0)}) = (2 - \frac{1}{2} \cdot 4, 3 - \frac{1}{2} \cdot 4) = (0,1)$.

Another step. We have $\nabla f(\mathbf{x}^{(1)}) = (-4,4)$, and $\phi_1(t) = f(4t, 1-4t)$. The derivative $\phi_1'(t)$ simplifies to $320t - 32$, and (in exactly the same way as in the first step) the global minimizer of $\phi_1(t)$ is at $t_1 = \frac{1}{10}$.

We set $\mathbf{x}^{(2)} = \mathbf{x}^{(1)} - \frac{1}{10}\nabla f(\mathbf{x}^{(1)}) = (0 - \frac{1}{10} \cdot -4, 1 - \frac{1}{10} \cdot 4) = (\frac{2}{5}, \frac{3}{5})$.

If we continue this process, we get the sequence of points

$$\mathbf{x}^{(0)} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}, \mathbf{x}^{(1)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \mathbf{x}^{(2)} = \begin{bmatrix} 0.4 \\ 0.6 \end{bmatrix}, \mathbf{x}^{(3)} = \begin{bmatrix} 0 \\ 0.2 \end{bmatrix}, \mathbf{x}^{(4)} = \begin{bmatrix} 0.08 \\ 0.12 \end{bmatrix}, \dots$$

and the values of f at these points are correspondingly $10, 2, 0.4, 0.08, 0.016, \dots$.

We always keep getting closer and closer to $(0,0)$. Compared to Newton's method, this is slow (linear) convergence.

If you plot the points, you see another peculiar feature of this method. We are moving in a zigzag: consecutive steps are at right angles to each other.

2 Properties of steepest descent

We can prove that this observation always holds.

Theorem 2.1 (Theorem 3.2.3 in the textbook). *If $\mathbf{x}^{(k)}, \mathbf{x}^{(k+1)}, \mathbf{x}^{(k+2)}$ are consecutive iterations of steepest descent, then*

$$(\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}) \cdot (\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}) = 0.$$

Proof. The step $\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}$ is parallel to $\nabla f(\mathbf{x}^{(k)})$, and the next step $\mathbf{x}^{(k+2)} - \mathbf{x}^{(k+1)}$ is parallel to $\nabla f(\mathbf{x}^{(k+1)})$. So we want to prove that $\nabla f(\mathbf{x}^{(k)}) \cdot \nabla f(\mathbf{x}^{(k+1)}) = 0$.

Since $\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - t_k \nabla f(\mathbf{x}^{(k)})$, where t_k is the global minimizer of $\phi_k(t) = f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)}))$, in particular it is a critical point, so $\phi_k'(t_k) = 0$.

The theorem follows from here: we have

$$\phi_k'(t) = \nabla f(\mathbf{x}^{(k)} - t\nabla f(\mathbf{x}^{(k)})) \cdot (-\nabla f(\mathbf{x}^{(k)}))$$

by the chain rule, so when we set $t = t_k$, we get $\phi_k'(t_k) = -[\nabla f(\mathbf{x}^{(k+1)}) \cdot \nabla f(\mathbf{x}^{(k)})]$. We know this should be 0 because t_k is a critical point. □

Intuitively, once we pick the direction of steepest descent, we move in that direction for as long as this benefits us: until f starts increasing again. This means that when we stop, the direction we were moving at is useless to us, and the new direction of steepest descent should be orthogonal to the previous one.

The good side of steepest descent is that under some hypotheses, we can guarantee convergence no matter how bad our initial guess is. (Also, as an advantage over Newton's method, we are actually making some attempt to minimize rather than maximize.)

First of all, we have

Theorem 2.2 (Theorem 3.2.5 in the textbook)**.** *The sequence* $f(\mathbf{x}^{(0)}), f(\mathbf{x}^{(1)}), f(\mathbf{x}^{(2)}), \dots$ *is strictly decreasing until/unless it reaches a critical point.*

Proof. The sequence is decreasing because we choose t_k as a global minimizer at each step:

$$f(\mathbf{x}^{(k+1)}) = \phi_k(t_k) \le \phi_k()) = f(\mathbf{x}^{(k)}).$$

If equality holds, then because t_k is a global minimizer of ϕ_k, 0 is *also* a global minimizer of ϕ_k. Therefore $\phi_k'(0) = 0$, which gives us

$$\nabla f(\mathbf{x}^{(k)} + 0\nabla f(\mathbf{x}^{(k)})) \cdot [-\nabla f(\mathbf{x}^{(k)})] = 0 \implies \|\nabla f(\mathbf{x}^{(k)})\|^2 = 0.$$

This makes $\mathbf{x}^{(k)}$ a critical point of f. □

We can't guarantee that if we stop, we stop at a local minimizer, because as soon as we reach any critical point, we can't keep going. But usually we will be approaching a point, rather than stopping.

We can guarantee that the method works under the following assumptions:

Theorem 2.3 (Theorem 3.2.6/3.2.7 in the textbook)**.** *If f is coercive and has a unique critical point (which must then be a global minimizer) then the method of steepest descent finds it.*

Proof. In outline, the proof proceeds in three steps:

1. We show that a subsequence of the steepest descent sequence converges to a point $\mathbf{x}^*$.
2. We show that this $\mathbf{x}^*$ must be a critical point.

3. We show that actually, the whole steepest descent sequence must converge to $\mathbf{x}^*$.

Step 1 follows from applying the Bolzano–Weierstrass theorem. Every sequence in a closed and bounded set has a convergent subsequence. We know that the steepest descent sequence stays in a closed and bounded set because once we start from $\mathbf{x}^{(0)}$, we can only get to smaller values of f. So we stay in the set

$$\{\mathbf{x} \in \mathbb{R}^n : f(\mathbf{x}) \le f(\mathbf{x}^{(0)})\}$$

which is closed because it's defined by a $\le$ inequality, and bounded because f is coercive.

Step 2 is analogous to the way we proved the validity of Newton's method. If $\mathbf{x}^*$ were not a critical point, we could do a single step of steepest descent to get to a point $\mathbf{x}^{**} = \mathbf{x}^* - t\nabla f(\mathbf{x}^*)$ with $f(\mathbf{x}^{**}) < f(\mathbf{x}^*)$.

By continuity, if we have a sequence $\mathbf{y}^{(1)}, \mathbf{y}^{(2)}, \mathbf{y}^{(3)}, \dots$ (a subsequence of the steepest descent sequence) converging to $\mathbf{x}^*$, then we must also have

$$\lim_{k\to\infty} [\mathbf{y}^{(k)} - t_*\nabla f(\mathbf{y}^{(k)})] = \mathbf{x}^* - t_*\nabla f(\mathbf{x}^*) = \mathbf{x}^{**}.$$

But what is $\mathbf{y}^{(k)} - t_*\nabla f(\mathbf{y}^{(k)})$? It's an attempt at a steepest descent iteration from $\mathbf{y}^{(k)}$, but it's not a very good one: it uses the value t_* rather than whatever the best value of t to use for $\mathbf{y}^{(k)}$ is. So $f(\mathbf{y}^{(k)} - t_*\nabla f(\mathbf{y}^{(k)}))$ is not as good as the f-value we get after a proper step of steepest descent from $\mathbf{y}^{(k)}$, which in turn is not as good as $f(\mathbf{x}^*)$, the value we get after many steps of steepest descent.

This gives us the inequality

$$f(\mathbf{y}^{(k)} - t_*\nabla f(\mathbf{y}^{(k)})) \ge f(\mathbf{x}^*)$$

and, taking the limit as $k \to \infty$, it gives us $f(\mathbf{x}^{**}) \ge f(\mathbf{x}^*)$. But this is the opposite of our previous conclusion $f(\mathbf{x}^{**}) < f(\mathbf{x}^*)$! Contradition. So $\mathbf{x}^*$ must be a critical point.

Step 3 is also done by contradiction. Suppose that the steepest descent sequence $\mathbf{x}^{(k)}$ does not converge to $\mathbf{x}^*$. Then there is some $\epsilon > 0$ for which infinitely many terms of the sequence satisfy $|\mathbf{x}^{(k)} - \mathbf{x}^*| > \epsilon$.

Take all those terms. By the argument in step 1, they also contain a convergent subsequence, which has some limit $\mathbf{x}^?$ different from $\mathbf{x}^*$. By the argument in step 2, $\mathbf{x}^?$ must also be a critical point. But we assumed that f only had one critical point, so this is impossible. Therefore the entire steepest descent sequence converges to $\mathbf{x}^*$. □