Math 484: Nonlinear Programming[1] Mikhail Lavrov

Chapter 3, Lecture 4: More General Descent Methods

April 22, 2019 *University of Illinois at Urbana-Champaign*

1 What is a descent method?

Today, we work in the same setting as in the previous lecture: we have a function $f : \mathbb{R}^n \to \mathbb{R}$ with continuous first derivatives, and we want to minimize it under the assumption that working with multivariable functions is hard, but working with single-variable functions is easy. We want to improve on the flaws in the method of steepest descent, while keeping the good parts.

The method of steepest descent is a greedy strategy for minimizing a function. In that method, to minimize f, we:

- Pick a direction to go in where f decreases as quickly as possible.
- Go in that direction to the point where f would start increasing again if we went any further.

Today, we will consider more general descent methods, giving us more flexibility. Informally, we will want to:

- Pick a direction to go in where f is decreasing.
- Go in that direction far enough to notice, but not too far.

This is vague, so we will make what we want more precise. We will define some criteria that a descent method has to satisfy, then prove that it's always possible to satisfy these criteria.

2 Criteria for a descent method

Our general iteration step from a point $\mathbf{x}^{(k)}$ is to go to a point

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} + t_k \mathbf{p}^{(k)}$$

where $\mathbf{p}^{(k)}$ is some yet-to-be-determined direction, and t_k is some yet-to-be-determined step size. We could, of course, write $t_k \mathbf{p}^{(k)}$ down as a single vector. But we want to keep the two separate: we will usually first choose the direction $\mathbf{p}^{(k)}$, and then choose how far we want to go in that direction.

We assume that $\nabla f(\mathbf{x}^{(k)}) \neq \mathbf{0}$, because this is a first-order method; if $\nabla f(\mathbf{x}^{(k)}) = \mathbf{0}$, we don't know anything about where to go.

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

Criterion 1. We must have $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$.

This is a rather minimal requirement; we simply want to improve our position with every step. I mention it because the textbook refers to this criterion, but it will actually follow from the other criteria we introduce: you will never have to check this criterion separately.

Criterion 2. We must have $\mathbf{p}^{(k)} \cdot \nabla f(\mathbf{x}^{(k)}) < 0$.

We can characterize this criterion as asking for a "descent direction": a direction in which f is decreasing, at least initially.

This criterion will be our only condition on the direction $\mathbf{p}^{(k)}$; the other criteria will deal with the step size t_k. Today, we will not discuss how to choose a descent direction; we will have more to say about that later.

For the final two criteria, it will help to work with the one-variable function

$$\phi_k(t) = f(\mathbf{x}^{(k)} + t\mathbf{p}^{(k)})$$

that parametrizes the values of f going in the direction $\mathbf{p}^{(k)}$ from $\mathbf{x}^{(k)}$. (They can be phrased in terms of f entirely, which is what the textbook does, but they're really about ϕ_k.)

Criterion 2 can also be phrased in terms of ϕ_k: it says that $\phi_k'(0) < 0$.

Criterion 3, with parameter $\beta \in (0, 1)$. We must have

$$\phi_k'(t_k) \geq \beta \phi_k'(0).$$

Think of this requirement as saying "don't take steps that are too small". Since $\phi'(0) < 0$, we can always make a tiny bit of progress just by setting t_k to be some sufficiently small value. But if we do this, then at the next step, $\mathbf{x}^{(k+1)}$ will be approximately the same as $\mathbf{x}^{(k)}$, $\nabla f(\mathbf{x}^{(k+1)})$ will be approximately the same as $\nabla f(\mathbf{x}^{(k)})$, and our options for $\mathbf{p}^{(k+1)}$ will look approximately the same as for our options for $\mathbf{p}^{(k)}$ did. Nothing will have changed, so we'll have wasted a step.

Asking something like "choose t_k at least 0.01" wouldn't be helpful: absolute constants like 0.01 don't make sense for all functions. So instead, we ask that t_k is large enough that the picture has changed sufficiently: the rate of change of ϕ_k' at t_k is sufficiently different from what it was at 0.

Keep in mind when reading this criterion that $\phi_k'(0)$ is negative. So asking that $\phi_k'(t_k) \geq \beta\phi_k'(0)$ is asking that $\phi_k'(t_k)$ is *less* negative. Imagine that $\beta = \frac{1}{2}$ or so: then we're asking that the rate of decrease has *slowed down* by a factor of 2.

The value of β is a parameter we can tweak to tell our method how much we care about this criterion. It's useful to look at the two extreme values of β to get a feeling for it.

- If we set $\beta = 0$, we'd get back the method of steepest descent again, more or less: this criterion would tell us to keep going until we reach (or go past) a critical point of ϕ_k.
- If we set $\beta = 1$, this criterion would only ask "keep going until $\phi_k'(t_k) \geq \phi_k(0)$". So just $t_k = 0$ (and in many cases, any $t_k > 0$, as well) would work.

So a value of β between 0 and 1 is somewhere between these two extremes.

Finally, note that this criterion never tells us to stop. It can tell us "keep moving in the same direction, you haven't made enough progress yet" or else "you can stop if you like now". Criterion 4 will be the complementary rule that tells us how far is too far.

Criterion 4, with parameter $\alpha \in (0, \beta)$. We must have

$$\phi_k(t_k) \leq \phi_k(0) + \alpha t_k \phi_k'(0).$$

The intuition here is "only keep going for as long as you're getting at least a fraction of the promised payoff".

Imagine setting $\alpha = 1$. (This is not an allowed value of α, but it's useful for understanding the idea.) In that case, on the right-hand side of the inequality in Criterion 4, we'd get $\phi_k(0) + t_k \phi_k'(0)$, which is the linear approximation to ϕ_k at 0.

In practice, we expect the graph of $\phi_k(t_k)$ to lie above this linear approximation. (For example, that's what would happen if f were convex.) This means that as soon as we take $t > 0$, $\phi_k(t)$ is not as small as the linear approximation predicted: we're getting less benefit by increasing t.

Criterion 4 allows for some amount of pessimism. Based on $\phi_k'(0)$, we can predict some rate at which ϕ_k will decrease when we increase t. We multiply this rate by some $\alpha > 0$ that represents our pessimism (or our tolerance for small amounts of progress); instead of expecting a rate of change of $\phi_k'(0)$, we'll only expect a rate of change of $\alpha \phi_k'(0)$. Then Criterion 4 tells us that we've gone too far if the actual improvement falls short even of this expectation.

One consequence of Criterion 4 is that we definitely have $\phi_k(t_k) < \phi_k(0)$, and therefore $f(\mathbf{x}^{(k+1)}) < f(\mathbf{x}^{(k)})$, satisfying Criterion 1. This is because (by Criterion 2) $\phi_k'(0) < 0$, and no matter what α we multiply this by, Criterion 4 still expects *some* nonzero amount of improvement. So forget about Criterion 1.

Criterion 4 also helps us deal with situations where $\phi_k(t)$ looks like the graph of e^{-t}: it is always decreasing, but slower and slower over time, and never goes below a certain point. In this case, the method of steepest descent would choke, and go off to infinity, because it wants to minimize $\phi_k(t)$ at every step. But Criterion 4 would notice that at some point, the benefit of going further is not worth it.

3 Wolfe's theorem

Theorem 3.1 (Wolfe's theorem; 3.3.1 in the textbook)**.** *Let α, β be real numbers satisfying $0 < \alpha < \beta < 1$. Assume $f : \mathbb{R}^n \to \mathbb{R}$ has continuous ∇f and is bounded from below.*

Starting from a point $\mathbf{x}^{(k)}$, whenever $\mathbf{p}^{(k)}$ satisfies Criterion 2 (it is a descent direction), then there is a range $[a_k, b_k]$ (with $0 < a_k < b_k$) such that for any $t_k \in [a_k, b_k]$, all four criteria are satisfied.

Proof. We work exclusively with the one-variable function $\phi_k(t)$, since we're only trying to decide which values of t_k will work. By Criterion 2, we know that $\phi_k'(0) < 0$.

Relative to $\phi_k'(0)$, we have two rates of change to keep in mind. One is the "slow rate" $\alpha\phi_k'(0)$: Criterion 4 asks that on average, we should decrease at least as fast as the slow rate, and stop when we can't.

The other rate is the "fast rate" $\beta\phi_k'(0)$: Criterion 3 asks that we keep going for as long as we're decreasing faster than the fast rate. Keep in mind that both rates are negative, so $\beta\phi_k'(0) \leq \alpha\phi_k'(0)$.

One possible choce of $[a_k, b_k]$ that will work is the following.

- Choose b_k to be the smallest value of t in $(0, \infty)$ such that $\phi_k'(t) = \alpha\phi_k'(0)$.
- Choose a_k to be the largest value of t in $(0, b_k)$ such that $\phi_k'(t) = \beta\phi_k'(0)$.

This choice of b_k guarantees that for any $t \in [0, b_k]$, $\phi_k'(t) \leq \alpha\phi_k'(0)$: the rate of change is always faster than the slow rate. Intuitively, this means that he average change should always be better than what's predicted by the slow rate. Algebraically, if $t_k \in [0, b_k]$, then

$$\int_0^{t_k} \phi_k'(t) \leq \int_0^{t_k} \alpha\phi_k'(0) \implies \phi_k(t_k) - \phi_k(0) \leq \alpha t_k \phi_k'(0)$$

which is the statement of Criterion 4.

This choice of a_k guarantees that for any $t_k \in [a_k, b_k]$,

$$\alpha\phi_k'(0) \geq \phi_k'(t_k) \geq \beta\phi_k'(0)$$

and therefore in particular Criterion 3 is satisfied.

But we should check that it's possible to choose these values of a_k and b_k. Any statement like "choose the smallest value" requires two things: first, that the set of values is nonempty, and second, that it has a smallest element.

Let's start with b_k. First of all, there must be *some* values t such that $\phi_k'(t) = \alpha\phi_k'(0)$. This is because if $\phi_k'(t) < \alpha\phi_k'(0)$ for all $t > 0$, then ϕ_k would keep decreasing faster than the slow rate forever, and then it would not be bounded below.

Second, there must be a smallest value. This is because $\{t : t \geq 0 \text{ and } \phi_k'(t) = \alpha\phi_k'(0)\}$ is a closed set, so it has a point closest to 0. Moreover, that point is not zero, because $\phi_k'(0) \neq \alpha\phi_k'(0)$. So b_k exists and is positive.

A similar argument holds for a_k. First, there must be some values $t \in [0, b_k]$ such that $\phi_k'(t) = \beta\phi_k'(0)$ by the intermediate value theorem applied to ϕ_k': since ϕ_k' starts off at $\phi_k'(0)$ and ends at $\alpha\phi_k'(0)$, so it must pass through $\beta\phi_k'(0)$ at some point.

Second, there must be a largest value. This is because $\{t : 0 \leq t \leq b_k \text{ and } \phi_k'(t) = \beta\phi_k'(0)\}$ is a closed set, so it has a point closest to b_k. That point is not 0 or b_k, because $\phi_k'(0) \neq \beta\phi_k'(0) \neq \alpha\phi_k'(0)$. So a_k exists and $0 < a_k < b_k$.

So we have an interval $[a_k, b_k]$ where Criteria 3 and 4 are satisfied. Criterion 1 follows automatically, and we already assumed Criterion 2, so we are done. □

Next time, we will look at how to use such a descent method in practice.