2022-03-28

Nonlinear equations and optimization

For the next month or so, we will be discussing methods for solving nonlinear systems of equations and multivariate optimization problems. We will devote most of our attention to four related problem classes:

$$(1) \qquad\qquad f(x) = 0, \quad f : \mathbb{R}^n \to \mathbb{R}^n$$

$$(2) \qquad\qquad \min_x f(x), \quad f : \mathbb{R}^n \to \mathbb{R}$$

$$(3) \qquad\qquad \min_x \|f(x)\|^2, \quad f : \mathbb{R}^n \to \mathbb{R}^m$$

$$(4) \qquad\qquad f(x(s), s) = 0, \quad f : \mathbb{R}^n \times R \to \mathbb{R}^n$$

We treat these problems as a unified group because the solution methods employ many of the same techniques, and insights gained from one problem can be applied to another. For example:

- We can turn the nonlinear system problem (1) into a non-negative least squares problem (3) problem by observing $f(x) = 0$ iff $\|f(x)\|^2 = 0$.

- The nonlinear least squares problem is a special case of the more general unconstrained optimization problem (2). We consider it as a special case because we can apply ideas for solving *linear* least squares problem to the nonlinear case.

- For differentiable functions, the minima we seek in the optimization problem (2) must occur at points where the gradient is zero, also known as *stationary points* or *critical points*. We find these points by solving a system of nonlinear equations.

- We might introduce parameter dependence (as in (4)) to understand the physics of a problem or as a mechanism to "sneak up" on the solution to otherwise hard problems.

In general, we will look to an optimization formulation as a way of judging progress, even if we are solving nonlinear equations. But in constructing algorithms, we will often look at things from the perspective of solving nonlinear systems of equations. Whatever approach we use, the numerical linear algebra tools from the start of the semester will play a central role.

Questions What are some linear or quadratic examples of each of the classes of problems described above? How do we know how to solve these simpler problems using methods from earlier in the class?

The big ideas

While we will see many technical tricks in the next month, I claim two as fundamental:

Fixed point iterations All our nonlinear solvers will be *iterative*. We can write most as *fixed point iterations*

$$(5) \qquad\qquad x^{k+1} = G(x^k),$$

which we hope will converge to a fixed point, i.e. $x^* = G(x^*)$. We often approach convergence analysis through the *error iteration* relating the error $e^k = x^k - x^*$ at successive steps:

$$(6) \qquad\qquad e^{k+1} = G(x^* + e^k) - G(x^*).$$

We have already seen one example of this paradigm when we discussed stationary methods for solving linear systems and fixed point iterations in one dimension.

Model-based methods Most nonlinear problems are too hard to solve directly. On the other hand, we can *model* hard nonlinear problems by simpler (possibly linear) problems as a way of building iterative solvers. The most common tactic — but not the only one! — is to approximate the nonlinear function by a linear or quadratic function and apply all the things we know about linear algebra.

If there is a third over-arching theme, it is *understanding problem structure*, whether to get good initial guesses for iterations, to obtain convergence proofs for methods, or to understand whether a (possibly non-unique) solution to a nonlinear system of equations or optimization problem is the "right" solution for the task at hand.

Differential calculus: a refresher

We need a good foundation of multivariable differential calculus to construct iterations and to understand their convergence. While you should have this as background already, it is worth spending some time refreshing the concepts and the notation.

From $\mathbb{R}$ to $\mathbb{R}^n$

A lot of multivariable calculus involves applying concepts from calculus in one variable, one direction at a time. Suppose $f : \mathbb{R}^n \to \mathbb{R}^m$, and we want to understand the behavior of f near $x \in \mathbb{R}^n$. We reduce to a one-dimensional problem by looking at the behavior along a direction $0 \neq u \in \mathbb{R}^n$:

$$g(s) \equiv f(x + su).$$

The *directional derivative* of f at x in the direction u is

$$\frac{\partial f}{\partial u}(x) = g'(0) = \left.\frac{d}{ds}\right|_{s=0} f(x + su).$$

If we cannot compute directional derivatives explicitly, we may choose to estimate them by a *finite difference approximation*, e.g.

$$\frac{\partial f}{\partial u}(x) \approx \frac{f(x + hu) - f(x)}{h}$$

for sufficiently small h. If f is smooth enough, this formula has $O(h)$ error. The most frequently used directional derivatives are the derivatives in the directions of the standard basis functions $e_1, \ldots, e_n$; these are the partial derivatives $\partial f/\partial x_j$. We may also sometimes use the more compact notation $f_{i,j} \equiv \partial f_i/\partial x_j$.

We can also compute higher-order derivatives

$$\frac{\partial^k f}{\partial u^k}(x) = g^{(k)}(0) = \left.\frac{d^k}{ds^k}\right|_{s=0} f(x + su),$$

or we can compute mixed directional derivatives by differentiating $\partial f/\partial u$ in some new direction v. We say $f \in C^k(\Omega, \mathbb{R}^m)$ for some $\Omega \subset \mathbb{R}^n$ if all directional derivatives of f (pure or mixed) up to order k exist and are

continuous in Ω; or, equivalently, if all the partials up to order k exist and are continuous in Ω. Sometimes the domain Ω is clear from context; in this case, we will simply say that f "is C^k." We say a function is C^0 if it is continuous.

If there are $k + 1$ continuous directional derivatives around x, we have the Taylor expansion

$$f(x + su) = \sum_{j=0}^{k} \frac{g^{(j)}(0)}{j!} s^j + \frac{g^{(k+1)}(\xi)}{(k+1)!} s^{k+1}$$

$$= \sum_{j=0}^{k} \frac{1}{j!} \frac{\partial^j f}{\partial u^j}(x) s^j + \frac{1}{(k+1)!} \frac{\partial^{k+1} f}{\partial u^{k+1}}(x + \xi u) s^{k+1}$$

where $0 \le \xi \le s$ is some intermediate point.

Questions   If $f : \mathbb{R} \to \mathbb{R}^m$ is twice differentiable, then

$$\|[f(0) + f'(0)s] - f(s)\| \le \frac{s^2}{2} \left(\max_{0 \le \xi \le s} \|f''(\xi)\| \right).$$

Why is this true? You can stick to the 2-norm if you want, though it is true more generally. It may be useful to use the fact that in general $\|v\| = \max_{\|u^*\|=1} u^* v$.

Derivatives and approximation

The function f is *differentiable* at x if there is a good affine (constant plus linear) approximation

$$f(x + z) = f(x) + f'(x)z + o(\|z\|),$$

where the *Jacobian* $f'(x)$ (also writen $J(x)$ or $\partial f / \partial x$) is the $m \times n$ matrix whose (i, j) entry is the partial derivative $f_{i,j} = \partial f_i / \partial x_j$. If f is differentiable, the Jacobian matrix maps directions to directional derivatives, i.e.

$$\frac{\partial f}{\partial u}(x) = f'(x)u.$$

If f is C^1 in some open neighborhood of x, it is automatically differentiable. There are functions with directional derivatives that are not differentiable,

but we will usually restrict our attention to C^1 functions if we use differentiability at all.

When multivariable calculus is taught to students without linear algebra as a prerequisite or co-requisite, the chain rule sometimes seems bizarre and difficult to remember. But once you think of derivatives as being about affine approximation, it becomes much simpler. Suppose $h = f \circ g$ where $g : \mathbb{R}^n \to \mathbb{R}^m$ and $f : \mathbb{R}^m \to \mathbb{R}^p$. Let $y = g(x)$, and consider first order approximations of f and g at y and x, respectively:

$$f(y + z) = f(y) + f'(y)z + o(\|z\|)$$
$$g(x + w) = g(x) + g'(x)w + o(\|w\|)$$

Then letting $z = g(x + w) - g(x) = g'(x)w + o(\|w\|)$, we have

$$h(x + w) = f(y) + f'(y)(g'(x)w + o(\|w\|) + o(\|z\|)$$
$$= f(y) + f'(y)g'(x)w + o(\|w\|)$$

Thus, we have $h'(x) = f'(y)g'(x)$; that is, the derivative of the composition is the composition of the derivatives.

A nest of notations

A nice notational convention we have seen before, sometimes called *variational* notation (as in "calculus of variations") is to write a relation between a first order change to f and to x. If f is differentiable at x, we write this as

$$\delta f = f'(x)\, \delta x$$

where δ should be interpreted as "first order change in." In introductory calculus classes, this is sometimes called a *total derivative* or *total differential*, though there one usually uses d rather than δ. There is a good reason for using δ in variational calculus, though, so that is typically what I do.

I like variational notation because I find it more compact than many of the alternatives. For example, if f and g are both differentiable maps from $\mathbb{R}^n$ to $\mathbb{R}^m$ and $h = f^T g$, then I make fewer mistakes writing

$$\delta h = (\delta f)^T g + f^T (\delta g), \quad \delta f = f'(x)\delta x, \quad \delta g = g'(x)\delta x$$

than when I write

$$h'(x) = g^T f'(x) + f^T g'(x)$$

even though the the two are exactly the same. We could also write partial derivatives using indicial notation, e.g.

$$h_{,k} = \sum_i (g_i f_{i,k} + g_{i,k} f_i).$$

Similarly, I like to write the chain rule for $h = f \circ g$ where composition makes sense as

$$\delta h = f'(g(x))\delta g, \quad \delta g = g'(x)\delta x.$$

But you could also write

$$h'(x) = f'(g(x))g'(x)$$

or

$$h_{i,k} = \sum_j f_{i,j}(g(x))g_{j,k}(x).$$

I favor variational notation, but switch to alternate notations when it seems to simplify life (e.g. I often switch to indicial notation if I'm working with computational mechanics). You may use any reasonably sensible notation you want in your homework and projects, but should be aware that there is more than one notation out there.

Lipschitz functions

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is *Lipschitz* with constant M on $\Omega \subset \mathbb{R}^n$ if

$$\forall x, y \in \Omega, \quad \|f(x) - f(y)\| \le M\|x - y\|.$$

Not every continuous function is Lipschitz; but if Ω is bounded and closed[1], then any function $f \in C^1(\Omega, \mathbb{R}^m)$ is Lipschitz with constant $M = \max_{x \in \Omega} \|f'(x)\|$.

Lipschitz constants will come up in several contexts when discussing convergence of iterations. For example, if $G : \Omega \to \Omega$ is Lipschitz with some constant less than one on Ω, we call it a *contraction mapping*, and we can show that fixed point iterations with G will converge to a unique fixed point in Ω. Lipschitz functions also give us a way to reason about approximation quality; for example, if $f'(x)$ is Lipschitz with constant M on Ω containing

---

[1] A compact set, for those of you who have taken some analysis

x, then we can tighten the usual asymptotic statement about linear approximation of f: if the line segment from x to $x + z$ lives in Ω, then

$$f(x + z) = f(x) + f'(x)z + e(z), \quad \|e(z)\| \le \frac{M}{2}\|z\|^2.$$

This also gives us a way to control the error in a finite difference approximation of $\partial f / \partial u$, for example.

Questions

- Is $x \mapsto \sqrt{x}$ Lipschitz on $(0, 1)$? On $(1, \infty)$? If so, what are the Lipschitz constants?

- Show that $x \mapsto |x|$ is Lipschitz on $\mathbb{R}$ with Lipschitz constant 1.

Quadratics and optimization

We now consider the case where $f : \mathbb{R}^n \to \mathbb{R}$. If f is C^1 on a neighborhood of x, the derivative $f'(x)$ is a row vector, and we have

$$f(x + z) = f(x) + f'(x)z + o(\|z\|).$$

The *gradient* $\nabla f(x) = f'(x)$ points in the direction of steepest ascent for the affine approximation:

$$f(x + su) = f(x) + f'(x)u \le f(x) + \|f'(x)\|\|z\|$$

with equality iff $z \propto \nabla f(x)$. Note that the gradient and the derivative are *not the same* – one is a row vector, the other a column vector!

If $f'(x)$ is nonzero, there is always an ascent direction $(\nabla f(x))$ and a descent direction $(-\nabla f(x))$ for f starting at x. Therefore, if f is C^1 then any minimum or maximum must be a *stationary point* or *critical point* where $f'(x) = 0$; equivalently, we could say a stationary point is where $\nabla f(x) = 0$ or where every directional derivative is zero. This fact is sometimes known as the *first derivative test.*

If f is a C^2 function, we can write a *second-order Taylor series*

$$f(x + z) = f(x) + f'(x)z + \frac{1}{2}z^T H z + o(\|z\|^2)$$

where H is the symmetric *Hessian matrix* whose (i, j) entry is the mixed partial $f_{,ij}$. We note in passing that if $f \in C^3$, or even if $f \in C^2$ and the second derivatives of f are Lipschitz, then we have the stronger statement that the error term in the expansion is $O(\|z\|^3)$.

If x is a stationary point then the first-order term in this expansion drops out, leaving us with

$$f(x + z) = f(x) + \frac{1}{2}z^T H z + o(\|z\|^2).$$

The function has a strong local minimum or maximum at x if the quadratic part does, i.e. if H is positive definite or negative definite, respectively. If H is strongly indefinite, with both positive and negative eigenvalues, then x is a saddle point. This collection of facts is sometimes known as the *second derivative test*.

Questions

- Consider the function

$$\rho(x, y) = \frac{\alpha x^2 + 2\beta xy + \gamma y^2}{x^2 + y^2}.$$

  What equation characterizes the stationary points?

- Argue that the Hessian of ρ defined above is nowhere positive definite.