Bonus Lecture: Solving Systems of Equations

Chris Conlon

September 6, 2020

Grad IO

Basic Setup

Often we are interested in solving a problem like this:

Root Finding $f(x) = 0$

Optimization $\arg\min_x f(x)$.

These problems are related because we find the minimum by setting: $f'(x) = 0$

Root Finding

Newton's Method for Root Finding

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot (x - x_0)^2 + o_p(3)$$

Suppose we wanted to find a root of the equation where $f(x^*) = 0$ and solve for x:

$$0 = f(x_0) + f'(x_0) \cdot (x - x_0)$$
$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

This gives us an iterative scheme to find x^*:

1. Start with some x_k. Calculate $f(x_k), f'(x_k)$
2. Update using $x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}$
3. Stop when $|x_{k+1} - x_k| < \epsilon_{tol}$.

Halley's Method for Root Finding

Consider the Taylor series for $f(x)$ approximated around $f(x_0)$:

$$f(x) \approx f(x_0) + f'(x_0) \cdot (x - x_0) + f''(x_0) \cdot (x - x_0)^2 + o_p(3)$$

Now let's consider the second-order approximation:

$$\begin{aligned} x_{n+1} &= x_n - \frac{2f\left(x_n\right) f'\left(x_n\right)}{2\left[f'\left(x_n\right)\right]^2 - f\left(x_n\right) f''\left(x_n\right)} = x_n - \frac{f\left(x_n\right)}{f'\left(x_n\right) - \frac{f(x_n)}{f'(x_n)} \frac{f''(x_n)}{2}} \\ &= x_n - \frac{f\left(x_n\right)}{f'\left(x_n\right)} \left[1 - \frac{f\left(x_n\right)}{f'\left(x_n\right)} \cdot \frac{f''\left(x_n\right)}{2f'\left(x_n\right)}\right]^{-1} \end{aligned}$$

- Last equation is useful because we only need to know $f(x_n)/f'(x_n)$ and $f''(x_n)/f'(x_n)$
- If we are lucky $f''(x_n)/f'(x_n)$ is easy to compute or ≈ 0 (Newton's method).

Root Finding: Convergence

How many iterations do we need? This is a tough question to answer.

- However we can consider convergence where $f(a) = 0$:

$$|x_{n+1} - a| \leq K_d * |x_n - a|^d$$

  - $d = 2$ (Newton's Method) quadratic convergence (we need $f'(x)$)
  - $d = 3$ (Halley's Method) cubic convergence (but we need $f''(x)$)

Root Finding: Fixed Points

Some (not all) equations can be written as $f(x) = x$ or $g(x) = 0 : f(x) - x = 0$.

- In this case we can iterate on the fixed point directly

$$x_{n+1} = f(x_n)$$

- Advantage: we only need to calculate $f(x)$.
- There need not be a unique solution to $f(x) = x$.
- But... this may or may not actually work.

Contraction Mapping Theorem/ Banach Fixed Point

Consider a set $D \subset \mathbb{R}^n$ and a function $f : D \rightarrow \mathbb{R}^n$. Assume

1. D is closed (i.e., it contains all limit points of sequences in D)
2. $x \in D \Longrightarrow f(x) \in D$
3. The mapping g is a contraction on D : There exists $q < 1$ such that

$$\forall x, y \in D : \quad \|f(x) - f(y)\| \leq q\|x - y\|$$

Then

1. There exists a unique $x^* \in D$ with $f\left(x^*\right) = x^*$
2. For any $x^{(0)} \in D$ the fixed point iterates given by $x^{(k+1)} := f\left(x^{(k)}\right)$ converge to x^* as $k \rightarrow \infty$
3. $x^{(k)}$ satisfies the a-priori error estimate $\left\|x^{(k)} - x^*\right\| \leq \frac{q^k}{1-q}\left\|x^{(1)} - x^{(0)}\right\|$
4. $x^{(k)}$ satisfies the a-posteriori error estimate $\left\|x^{(k)} - x^*\right\| \leq \frac{q}{1-q}\left\|x^{(k)} - x^{(k-1)}\right\|$

Some notes

- Not every fixed point relationship is a contraction.
- Iterating on $x_{n+1} = f(x_n)$ will not always lead to $f(x) = x$ or $g(x) = 0$.
- Convergence rate of fixed point iteration is slow or $q-$linear.
- When q is small this will be faster.
- q is sometimes called modulus of contraction mapping.
- A key example of a contraction: value function iteration!

Accelerated Fixed Points: Secant Method

Start with Newton's method and use the finite difference approximation

$$f'\left(x_{n-1}\right) \approx \frac{f\left(x_{n-1}\right) - f\left(x_{n-2}\right)}{x_{n-1} - x_{n-2}}$$

$$x_n = x_{n-1} - f\left(x_{n-1}\right) \frac{x_{n-1} - x_{n-2}}{f\left(x_{n-1}\right) - f\left(x_{n-2}\right)}$$

- This doesn't have the actual $f'(x_n)$ so it isn't quadratically convergent
- Instead is is superlinear with rate $q = \frac{1+\sqrt{5}}{2} = 1.618 < 2$ (Golden Ratio)
- Faster than fixed-point iteration but doesn't require computing $f'(x_n)$.
- Idea: can use past iterations to approximate derivatives and accelerate fixed points.

Accelerated Fixed Points: Anderson (1965) Mixing

Define the residual $r(x_n) = f(x_n) - x_n$. Find weights on previous k residuals:

$$\widehat{\alpha^n} = \arg\min_{\alpha} \left\| \sum_{k=0}^{m} \alpha_k^n \cdot r_{n-k} \right\| \text{ subject to } \sum_{k=0}^{m} \alpha_k^n = 1$$

$$x_{n+1} = (1-\lambda) \sum_{j=0}^{m} \widehat{\alpha_k^n} \cdot x_{n-k} + \lambda \sum_{j=0}^{m} \widehat{\alpha_k^n} \cdot f\left(x_{n-k}\right)$$

- Convex combination of weighted average of: lagged x_{n-k} and lagged $f(x_{n-k})$.
- Variants on this are known as Anderson Mixing or Anderson Acceleration.

Accelerated Fixed Points: SQUAREM (Varadhan and Roland 2008)

Define the residual $r(x_n) = f(x_n) - x_n$ and $v(x_n) = f \circ f\,(x_n) - f\,(x_n)$.

$$\begin{aligned} x_{n+1} = & x_n && -2s\left[f\,(x_n) - x_n\right] && +s^2\left[f \circ f\,(x_n) - 2f\,(x_n) + x_n\right] \\ = & x_n && -2sr && +s^2(v - r) \end{aligned}$$

Three versions of stepsize:

$$s_1 = \frac{r^t r}{r^t(v-r)}, \quad s_2 = \frac{r^t(v-r)}{(v-r)^t(v-r)}, \quad s_3 = -\sqrt{\frac{r^t r}{(v-r)^t(v-r)}}$$

Idea: use two iterations to construct something more like the quadratic/Halley method.

Note: I am hand-waving, don't try to derive this.

Newton-Raphson for Minimization

We can re-write optimization as root finding;

- We want to know $\hat{\theta} = \arg\max_\theta \ell(\theta)$.
- Construct the FOCs $\frac{\partial \ell}{\partial \theta} = 0 \rightarrow$ and find the zeros.
- How? using Newton's method! Set $f(\theta) = \frac{\partial \ell}{\partial \theta}$

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta_k)\right]^{-1} \cdot \frac{\partial \ell}{\partial \theta}(\theta_k)$$

The SOC is that $\frac{\partial^2 \ell}{\partial \theta^2} > 0$. Ideally at all θ_k.
This is all for a single variable but the multivariate version is basically the same.

Newton's Method: Multivariate

Start with the objective $Q(\theta) = -l(\theta)$:

- Approximate $Q(\theta)$ around some initial guess θ_0 with a quadratic function
- Minimize the quadratic function (because that is easy) call that θ_1
- Update the approximation and repeat.

$$\theta_{k+1} = \theta_k - \left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1} \frac{\partial Q}{\partial\theta}(\theta_k)$$

- The equivalent SOC is that the Hessian Matrix is positive semi-definite (ideally at all θ).
- In that case the problem is globally convex and has a unique maximum that is easy to find.

Newton's Method

We can generalize to Quasi-Newton methods:

$$\theta_{k+1} = \theta_k - \lambda_k \underbrace{\left[\frac{\partial^2 Q}{\partial\theta\partial\theta'}\right]^{-1}}_{A_k} \frac{\partial Q}{\partial\theta}(\theta_k)$$

Two Choices:

- Step length λ_k
- Step direction $d_k = A_k \frac{\partial Q}{\partial\theta}(\theta_k)$
- Often rescale the direction to be unit length $\frac{d_k}{\|d_k\|}$.
- If we use A_k as the true Hessian and $\lambda_k = 1$ this is a full Newton step.

Newton's Method: Alternatives

Choices for A_k

- $A_k = I_k$ (Identity) is known as gradient descent or steepest descent
- BHHH. Specific to MLE. Exploits the Fisher Information.

$$A_k = \left[\frac{1}{N} \sum_{i=1}^{N} \frac{\partial \ln f}{\partial \theta} (\theta_k) \frac{\partial \ln f}{\partial \theta'} (\theta_k) \right]^{-1}$$

$$= -\mathbb{E}\left[\frac{\partial^2 \ln f}{\partial \theta \partial \theta'} (Z, \theta^*) \right] = \mathbb{E}\left[\frac{\partial \ln f}{\partial \theta} (Z, \theta^*) \frac{\partial \ln f}{\partial \theta'} (Z, \theta^*) \right]$$

- Alternatives SR1 and DFP rely on an initial estimate of the Hessian matrix and then approximate an update to A_k.
- Usually updating the Hessian is the costly step.
- Non invertible Hessians are bad news.

Extended Example: Binary Choice

Binary Choice: Overview

Many problems we are interested in look at discrete rather than continuous outcomes:

- Entering a Market/Opening a Store
- Working or a not
- Being married or not
- Exporting to another country or not
- Going to college or not
- Smoking or not
- etc.

Simplest Example: Flipping a Coin

Suppose we flip a coin which is yields heads ($Y = 1$) and tails ($Y = 0$). We want to estimate the probability p of heads:

$$Y_i = \begin{cases} 1 \text{ with probability } p \\ 0 \text{ with probability } 1 - p \end{cases}$$

We see some data $Y_1, \ldots, Y_N$ which are (i.i.d.)

We know that $Y_i \sim Bernoulli(p)$.

Simplest Example: Flipping a Coin

We can write the likelihood of N Bernoulli trials as

$$Pr(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N) = f(y_1, y_2, \ldots, y_N|p)$$

$$\begin{aligned} &= \prod_{i=1}^{N} p^{y_i}(1-p)^{1-y_i} \\ &= p^{\sum_{i=1}^{N} y_i}(1-p)^{N-\sum i=1^N y_i} \end{aligned}$$

And then take logs to get the log likelihood:

$$\ln f(y_1, y_2, \ldots, y_N|p) = \left(\sum_{i=1}^{N} y_i\right) \ln p + \left(N - \sum_{i=1}^{N} y_i\right)(1-p)$$

Simplest Example: Flipping a Coin

Differentiate the log-likelihood to find the maximum:

$$
\begin{aligned}
\ln f(y_1, y_2, \ldots, y_N | p) &= \left(\sum_{i=1}^{N} y_i\right) \ln p + \left(N - \sum_{i=1}^{N} y_i\right) \ln(1-p) \\
\rightarrow 0 &= \frac{1}{\hat{p}} \left(\sum_{i=1}^{N} y_i\right) + \frac{-1}{1-\hat{p}} \left(N - \sum_{i=1}^{N} y_i\right) \\
\frac{\hat{p}}{1-\hat{p}} &= \frac{\sum_{i=1}^{N} y_i}{N - \sum_{i=1}^{N} y_i} = \frac{\overline{Y}}{1-\overline{Y}} \\
\hat{p}^{MLE} &= \overline{Y}
\end{aligned}
$$

That was a lot of work to get the obvious answer: fraction of heads.

More Complicated Example: Adding Covariates

We probably are interested in more complicated cases where p is not the same for all observations but rather $p(X)$ depends on some covariates. Here is an example from the Boston HMDA Dataset:

- 2380 observations from 1990 in the greater Boston area.
- Data on: individual Characteristics, Property Characteristics, Loan Denial/Acceptance (1/0).
- Mortgage Application process circa 1990-1991:
  - Go to bank
  - Fill out an application (personal+financial info)
  - Meet with loan officer
  - Loan officer makes decision
    - Legally in race blind way (discrimination is illegal but rampant)
    - Wants to maximize profits (ie: loan to people who don't end up defeaulting!)