Chapter 2: Optimization

The Role of Optimization in Inference

For many statistical inference problems, the estimator of the parameter of interest, is defined as the point at which some function involving the parameter along with realizations of the random variable achieves an optimum. (max or min)

Notation:

- Let Y be r.v. of interest
- Let $(y_1, \ldots, y_n)$ be realizations of Y
- let θ be the variable denoting parameter with $\theta \in \Theta$
- let θ_* be the fixed true value of θ
- Let $s(\theta)$ be a real valued function of θ, $s: \Theta \to \mathbb{R}$.
- Let $\hat{\theta}$ denote the estimator of θ.

We have two type of problems.

(1) Minimization

$$\hat{\theta} = \arg\min_{\theta \in \Theta} s(\theta)$$

Ex/ Minimum residual

(2) Maximization

$$\hat{\theta} = \arg\max_{\theta \in \Theta} s(\theta)$$

Ex Maximum Liklihood Estimation

The ideal scenario for the above opt. prob arises when $s(\Theta)$ is a smooth "well behaved" function w/ nice prop. (e.g. $s(\theta)$ is diff, bounded, convex, etc.) and has a unique optimum that can be expressed in closed form.

Q: Where does comp. stats. intervene?

A: Many statistical analyses do not admit a closed form expression for $\hat{\Theta}$ and it often happens that $s(\Theta)$ does not behave.

(many local optima, non differentiable, non continuous, Θ could be constrained.).

See handout

Estimation by Minimizing Residuals

In many applications we can express the exp. val of a r.v. as a function of a parameter

$$E[Y] = f(\Theta_*).$$

So, if we have observations $y_1, \ldots, y_n$ on Y then a reasonable estimator of Θ_* is a value $\hat{\Theta}$ that minimizes the residuals

$$r_i(\Theta) = y_i - f(\Theta)$$

over all possible choices of Θ.

This makes sense since we expect the y_i to be close to $f(\Theta_*)$.

There are several ways we could reasonably "minimize the residuals", but in general we seek to minimize some norm of $\vec{r}(\theta) = \begin{bmatrix} r_1(\theta) \\ \vdots \\ r_n(\theta) \end{bmatrix}$ and so our opt. prob becomes

doesn't have to be Lp norm.

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \|r(\theta)\| = \underset{\theta \in \Theta}{\operatorname{argmin}}\ S_p(\theta)$$

where

$$S_p(\theta) = \sum_{i=1}^{n} |y_i - f(\theta)|^p \quad \text{the } L_p \text{ norm.}$$

in this case $\hat{\theta}$ is called the <u>Lp estimator</u>.

- if $p=2$, we have the least squares estimator.

Note:

More generally, let $\rho(\cdot)$ be some nonnegative function, then minimizing

$$S_\rho(\theta) = \sum_{i=1}^{n} \rho(y_i - f(\theta))$$

yields the so-called M-estimator, which are often more robust than standard least-squares (or Lp) and are usually used to try and reduce the effect of outliers.

Now,

The question is how do we optimize this type of problem?

We begin by discussing some general optimization techniques.

Back to Chpt 2.

Many functions can be optimized analytically

Ex/ Find min $f(x) = (x-1)^2 = x^2 - 2x + 1$

Then

$$f'(x) = 2x - 2 = 2(x-1)$$

and $f'(x) = 0$ has solution $x = 1$

also

$$f'' = 2 \quad \therefore \quad x = 1 \text{ is a minimum}$$

However many functions cannot: ex maximize $g(x) = \frac{\log x}{1+x}$

$g(x) = 1 + 1/x - \log x$, and $g'(x) = 0$ has no analytic solution.

What to do? What to do?

Assumptions

1. want to optimize $g(x)$ w.r.t. x (p-dim)
2. assume maximization since it is eq. to minimizing its negative
3. g is smooth & diff.
   discrete opt discussed in Ch. 3.

So we have the root finding problem

$$g'(\vec{x}) = \vec{0} \quad \text{that maximizes}$$

If the system is linear → many Linear programming techniques which are gaurenteed (simplex etc) & not discussed.

However often → we encounter nonlinear systems. and so our methods will be numerical & iterative. start w/ some x_0 & find x_t $t = 1, 2, \ldots$ until done.

2 questions ① where to start
② when are we done.

Sec 2.1
pg 20

Univariate Problems

How to maximize

$$g(x) = \frac{\log x}{1+x} \Rightarrow g'(x) = 0$$

want

where to start

where to start.

If we graph g we see (pg 21) that the max is between 3 & 4 and so we will use this information ~~for~~ in an iterative procedure starting point

Bisection Method

"A nice example of an iterative procedure."

If g' is continuous on $[a_0, b_0]$ & $g'(a_0)g'(b_0) \leq 0$ then by IVT $\exists\, x^* \in [a_0, b_0] \ni g'(x^*) = 0$. and is a local optimum.

The basic idea is to shrink intervals $[a_0, b_0] \supseteq [a_1, b_1] \supset [a_2, b_2]$.

pseudo Algorithm

- Let $x^{(0)} = (a_0 + b_0)/2$
- $[a_{t+1}, b_{t+1}] = \begin{cases} [a_t, x^{(t)}] & \text{if } g'(a_t)g'(x^{(t)}) \leq 0 \\ [x^{(t)}, b_t] & \text{if } g'(a_t)g'(x^{(t)}) > 0 \end{cases}$
- $x^{(t+1)} = (a_{t+1}, b_{t+1})/2$

Ex/pg. 22 $[1, 5]$ & $x^{(0)} = 3$ goes to $x^* = 3.5912$ after 19 iterations.

Down side (only find 1 root in $[a_0, b_0]$)

Question 2: When to Stop.

We hope that $x^{(t)} \to x^*$. However there is no ~~g~~ gaurantee of this or even that $x^{(t)}$ will converge. And we don't want our procedure to run indefinitely so we require a stopping rule

Usually has 2 parts
1. convergence criteria (success?)
2. failure rule.

Convergence criteria

Want a rule that can be checked at each iteration, and once met $x^{(t+1)}$ is taken as solution.

flat curve
- The proximity of $g'(x^{(t+1)})$ to 0 is not very reliable since allow large Δ between $x^{(t)}$ & $x^{(t+1)}$.
- Usually based on small change between $x^{(t)}$ & $x^{(t+1)}$ (indicating g' close to 0).

Absolute Convergence

$$|x^{(t+1)} - x^{(t)}| < \varepsilon \quad \text{for some acceptable } \varepsilon > 0.$$

Ex/ Bisection

$$b_t - a_t = 2^{-t}(b_0 - a_0)$$

$\therefore$ If we require $|x^{(t)} - x^*| < \delta$ we can stop when $|x^{(t+1)} - x^{(t)}| = 2^{-(t+1)}(b_0 - a_0) < \delta$.

or after $t > \log_2\left(\frac{b_0 - a_0}{\delta}\right) - 1$ iterations.

This means that to increase our precision by 1 decimal place we need

$$t > \log_2\left(\frac{b_0 - a_0}{(\delta/10)}\right) - 1 = \log_2\left(\frac{b_0 - a_0}{\delta}\right) - 1 + \log_2 10$$

→ we must do $\log_2 10$ or ~ 3.3 more iterations.

Relative convergence

want % precision

$$\frac{|x^{(t+1)} - x^{(t)}|}{|x^{(t)}|} < \varepsilon$$

Also Bisection Method is gauranteed to converge to a root of g' since

$$\lim_{t\to\infty} a_t = \lim_{t\to\infty} b_t = x^{\infty} \quad \text{and} \quad g'(a_t)\, g'(b_t) \leq 0$$

implies by continuity that $\left[g'(x^{\infty})\right]^2 \leq 0 \Rightarrow g'(x^{\infty}) = 0$.

However it can be slow.

Further examples & discussion on pg 23-24.

(2) Failure rule.

(1) Stop after N iterations
(2) Stop if convergence measures fail to decrease. or cycle.

Sec 2.1.1 pg 24 **Newton's method** (Newton-Raphson iteration)

- extremely fast.

Suppose g' is cont. diff and $g''(x^*) \neq 0$.
then we approx $g'(x^*)$ by the linear Taylor expan. (see Sec 1.2)

$$0 = g'(x^*) \approx g'(x^{(t)}) + (x^* - x^{(t)})\, g''(x^{(t)})$$

the tangent line at $x^{(t)}$. $\therefore$ we approx the root by solving $0 = g'(x^{(t)}) + (x^* - x^{(t)})\, g''(x^{(t)})$ for x^*.

&

$$x^* \approx x^{(t+1)} \equiv x^{(t)} - \underbrace{g'(x^{(t)}) \big/ g''(x^{(t)})}_{\text{This is a refinement } h(t).}$$

Can converge quickly - pg 25

Ex/ $g(x) = \frac{\log x}{1+x}$ $\quad g'(x) = \frac{1 + 1/x - \log x}{(1+x)^2}$

$$h(x) = \frac{(x^{(t)}+1)(1 + 1/x^{(t)} - \log x^{(t)})}{3 + 4/x^{(t)} + 1/(x^{(t)})^2 - 2\log x^{(t)}}.$$

and w/ $x^{(0)} = 3$ $\quad x^4 = 3.59112$. (v/s 19)

Downside: Unlike Bisection $\Rightarrow$ not gaurenteed to converge. (pg 26).

However there is a region about x^* for which it will. See proof pg 26-27.

So $\Rightarrow$ you need a good starting guess.

2.1.1.1 Convergence order

We saw that Newton's method converged to our solution for $g'(x)=0$ more quickly than the Bisection Method.

In general how can we discuss the speed of a root-finding approach (or any other similar algorithm).
- need order of convergence.

Let $x^{(t)} - x^* = \varepsilon^{(t)}$, then a method has convergence order β if

① $\lim_{t\to\infty} \varepsilon^{(t)} = 0$

② $\lim_{t\to\infty} \frac{|\varepsilon^{(t+1)}|}{|\varepsilon^{(t)}|^\beta} = c$

for some $c \neq 0$ and $\beta > 0$.

- Higher orders of convergence are better in the sense that the answer is found more quickly, but often at the expense of robustness.

For Newton's method:

$$0 = g'(x^*) = g'(x^{(t)}) + (x^* - x^{(t)})g''(x^{(t)}) + (x^* - x^{(t)})^2 g'''(q)/2$$

for $q \in$ between $x^{(t)}$ & x^*.

$$\underbrace{x^{(t)} + h^{(t)}}_{x^{(t+1)}} - x^* = (x^* - x^{(t)})^2 \, g'''(q) / 2g''(x^{(t)})$$

$\therefore$

$$\varepsilon^{(t+1)} = (\varepsilon^{(t)})^2 \frac{g'''(q)}{2g''(x^{(t)})}$$

$$\therefore \quad \frac{\varepsilon^{(t+1)}}{(\varepsilon^{(t)})^2} = \frac{g'''(\xi)}{2g''(x^{(t)})}$$

Where $g'''(\xi) \longrightarrow g'''(x^*)$ and

$g''(x^{(t)}) \longrightarrow g''(x^{(t)})$

$\therefore$ Letting $c = \left| \dfrac{g'''(x^*)}{2g''(x^*)} \right|$

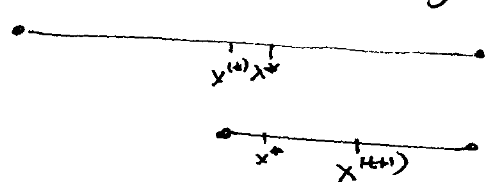
we see that $\displaystyle\lim_{t\to\infty} \frac{|\varepsilon^{(t+1)}|}{|(\varepsilon^{(t)})^2|} = c$

and Newton's method has convergence order $\beta = 2$.

What about Bisection Method?

① $\displaystyle\lim_{t\to\infty} |\varepsilon^{(t)}| \Rightarrow 0$ (since distance always halved)

② However $|x^{(t)} - x^*|$ may not shrink at each interval



$\therefore \displaystyle\lim_{t\to\infty} \frac{|\varepsilon^{(t+1)}|}{|\varepsilon^{(t)}|^\beta}$ may not exist.

$\therefore$ Does not have a formal convergence order (similiar to linear though).

But it is more robust.

Sec 1.4
pg 9

Maximum Likelihood Estimation Review

If $X_1, \dots, X_n$ are i.i.d w/ density $f(x|\theta)$ for $\theta = (\theta_1, \cdots \theta_p)$ then the joint likelihood function is

$$L(\theta) = \prod_{i=1}^{n} f(x_i|\theta).$$

If $x_1, \cdots, x_n$ are observed data, then the parameters θ most likely to have caused $x_1, \dots, x_n$ constitute the maximum likelihood estimate of θ and is the $\hat{\theta}$ = "function of data" that maximizes $L(\theta)$.

Typically it is easier to work w/

$$\ell(\theta) = \log L(\theta)$$

the log likelihood function since it has the same maximum as $L(\theta)$ and we can ignore additive constants.

So maximizing $L(\theta)$ is eq. to solving $\ell'(\theta) = 0$.
where

$$\ell'(\theta) = \left(\frac{d\ell(\theta)}{d\theta_1}, \frac{d\ell(\theta)}{d\theta_2}, \dots, \frac{d\ell(\theta)}{d\theta_p} \right)$$

is the score function and satisfies $E[\ell'(\theta)] = 0$.

Ex/ $N(\mu, \sigma^2)$ $\theta = (\mu, \sigma^2)$ $f(x|\theta) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

$$\ell(\theta|x) = -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2}$$

and so the score function is

$$\ell'(\theta) = \left(\frac{\partial \ell}{\partial \mu}, \frac{\partial \ell}{\partial \sigma^2}\right) = \left(\frac{x-\mu}{\sigma^2}, -\frac{1}{2\sigma^2} + \frac{(x-\mu)^2}{2\sigma^4}\right)$$

and

$$E[\ell'(\theta)] = (0, 0) \quad \text{since } E[(x-\mu)^2] = \sigma^2$$

Also let $\ell''(\theta)$ be the $p \times p$ matrix with $\left[\frac{\partial^2 \ell(\theta)}{\partial\theta_i \partial\theta_j}\right]_{ij}$ as its (i,j)th entry.

Then

$$I(\theta) = -E[\ell''(\theta)] = E[\ell'(\theta)\ell'(\theta)^T]$$

is called the Fisher information matrix.

$\ell''(\theta)$ is sometimes called the observed Fisher information matrix and is useful because it can always be calc & is a good approx to $I(\theta)$.

Ex/ $\ell''(\theta) = \begin{bmatrix} \frac{d^2\ell}{d\mu^2} & \frac{d^2\ell}{d\mu d\sigma^2} \\ \frac{d^2\ell}{d\mu d\sigma^2} & \frac{d^2\ell}{d(\sigma^2)^2} \end{bmatrix}$

$$= \begin{bmatrix} -\frac{1}{\sigma^2} & -\frac{(x-\mu)}{\sigma^4} \\ -\frac{(x-\mu)}{\sigma^4} & \frac{1}{2\sigma^4} - \frac{(x-\mu)^2}{\sigma^6} \end{bmatrix}$$

and so the Fisher information matrix is

$$I(\theta) = \underset{\text{neg.}}{-}E[\ell''(\theta)] = -\begin{bmatrix} -\frac{1}{\sigma^2} & 0 \\ 0 & \frac{1}{2\sigma^4} - \frac{1}{\sigma^4} \end{bmatrix} = \begin{bmatrix} \frac{1}{\sigma^2} & 0 \\ 0 & 1/2\sigma^4 \end{bmatrix}.$$

Sec. 2.1.2 pg 28 <u>Fisher Scoring</u>

Consider applying Newtons method to a MLE problem. Here maximizing $L(\theta)$ is equivalant to solving $\ell'(\theta)=0$. and Newton's Method becomes

$$\theta^{(t+1)} = \theta^{(t)} - \frac{\ell'(\theta^{(t)})}{\ell''(\theta^{(t)})} \qquad (*)$$

Now we know that $-\ell''(\theta)$ is an approx for the Fisher information matrix $I(\theta)$ and so it is resonable to use this replacement in eq(*). yielding

<u>Fisher Scoring</u>

$$\theta^{(t+1)} = \theta^{(t)} + \ell'(\theta^{(t)})\, I(\theta^{(t)})^{-1}$$

Both methods have the same asymptotic properties, but often one is easier than the other.
- ex - difficult 2nd derivative, hard exp.

Generally, F.S. works better in the beginning to make rapid improvements, while N.M. works better for refinement near the end.

Sec 2.1.3 <u>Secant Method</u>

Newton's Method relies on the 2nd derivative $g''(x^{(t)})$. Which can be difficult to calculate.
However we can replace it by the discrete diff approx

$$\frac{g'(x^{(t)}) - g'(x^{(t-1)})}{x^{(t)} - x^{(t-1)}}$$

Resulting in

$$x^{(t+1)} = x^{(t)} - g'(x^{(t)}) \frac{x^{(t)} - x^{(t-1)}}{g'(x^{(t)}) - g'(x^{(t-1)})} \quad \text{for } t \geq 1.$$

which is the Secant method. (use secant line between $x^{(t)}$ & $x^{(t-1)}$ to get $x^{(t+1)}$)

This method requires two starting values. It will converge to x^* under conditions sim. to those for Newton's. and has convergence order $\beta = 1.62$. (See discussion on pages 29-30).

2.1.4 pg 30 **Fixed - Point Iteration**

A fixed point of a function G is a point $x \ni G(x) = x$.

For our root finding problems we want $G \ni$
$$g'(x) = 0 \iff G(x) = x.$$
Letting $G(x) = g'(x) + x$ we get the algorithm
$$x^{(t+1)} = x^{(t)} + g'(x^{(t)})$$

Note: Both Newton's & Secant Method are ex of F.P.I.

Convergence: The convergence of the algorithm requires that G be contractive on $[a, b]$ i.e.

1. $G(x) \in [a,b]$ whenever $x \in [a,b]$
2. $|G(x_1) - G(x_2)| \leq \lambda |x_1 - x_2| \; \forall \; x_1, x_2 \in [a,b]$ & some $\lambda \in [0, 1)$.

is called the Lipschitz condition

If G is contractive than the algorithm is gauranteed to converge to a unique f.p. x^* on $[a,b]$. and
$$|x^{(t)} - x^*| \leq \frac{\lambda^t}{1-\lambda} |x^{(1)} - x^{(0)}|.$$
and the order of convergence will be dependent on λ.

Convergence is not gauranteed, however, if $g''(x)$ is bounded and does not change sign on $[a,b]$, then we can rescale nonconvergent problems by choosing $\alpha \neq 0$ and letting

$$G(x) = \alpha g'(x) + x.$$

This works since $\alpha g'(x) = 0$ iff $g'(x) = 0$.

There are ways to carefully calculate α, however it is often easier to just try a few values. (Ex pg 26).

Sec 2.2 Multivariate Problems.

In a multivariate opt. problem we seek to max/min a real valued function g of a p-dim vector $x = (x_1, \dots, x_p)^T$.
At iteration t, $x^{(t)} = (x_1^{(t)}, \dots, x_p^{(t)})^T$

Many of the general principles still apply.
- iterative algorithms
- often take steps based on linearization of g' from Taylor series, secant approx, etc.
- convergence criteria are in the same spirit

Convergence criteria

Lp norms

Need $D(u,v)$, a distance measure for p-dim vectors.
Ex/ $D(u,v) = \sum_{i=1}^{p} |u_i - v_i|$ or $D(u,v) = \sqrt{\sum_{i=1}^{p} (u_i - v_i)^2}$
Then we form ab, rel convergence from.

$$D(x^{(t+1)}, x^{(t)}) < \varepsilon \quad \text{or} \quad \frac{D(x^{(t+1)}, x^{(t)})}{D(x^{(t)}, 0)} < \varepsilon$$