TOPIC: Biostat 778

DATE: 2005-10-27

FILE UNDER:

PAGE:

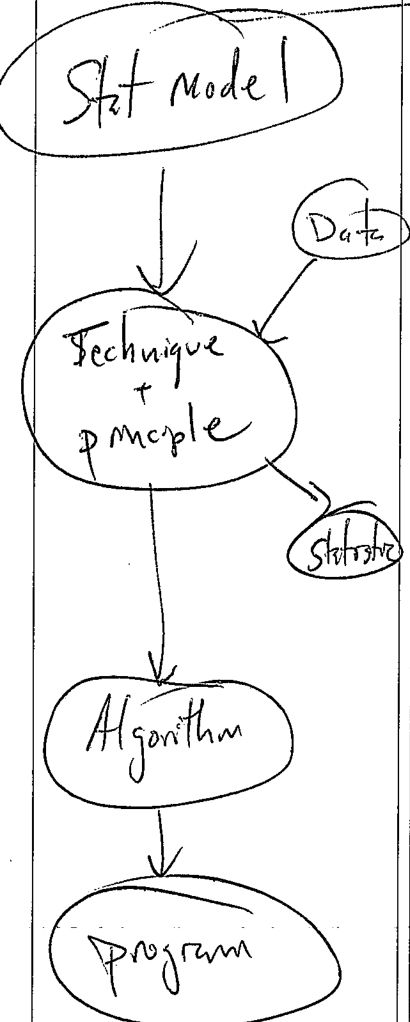
① ~~Solutions to non linear equations~~

~~Let $f: \mathbb{R} \to \mathbb{R}$. Solve $f(x) = 0$ for $x \in [a, b]$~~

~~- Bisection~~

~~- functional iteration~~

- Newton's method



~~We~~ Statistics is:

① Probability
② Linear Algebra
③ Optimization ⟸

We look at ③.

Generally, we want to maximize or minimize something.

⇒ max! likelihood
⇒ min! sum of squares

⇒ ~~max~~ $\max f = \min -f$
so don't worry about it.

This course is about

① maximizing a function
② integrating a function

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~Deterministic Algorithms~~

Entire Class/course

General idea!

1. It's difficult to ~~optimize~~ maximize f.
2. We can compute an approximation to f called g.
3. "Transfer" optimization to g and maximize g.
4. Iterate 2, 3

$\Rightarrow$ Instead of direct max, "transfer" to simpler function and iterate

$$f(b) - f(a) = f'(c)(b-a)$$

$$f(x_n - \ ) - f(x_\infty) = f'(c)(\cancel{x_{n-1}} - x_\infty)$$

$$f(x_{n-1}) = f'(c)(x_{n-1} - x_\infty)$$

~~Course outline~~

TOPIC:

DATE:

FILE UNDER:

PAGE:

Course outline

Maximizing

(*) Solving non linear equations (root finding)

$f(x) = 0$ for $f : \mathbb{R} \to \mathbb{R}$

$f(x) = 0$ for $f : \mathbb{R}^K \to \mathbb{R}$

$K = 2, 3, \ldots, N$ (~~[illegible]~~)

(*) General optimization routines

Given $f : \mathbb{R}^K \to \mathbb{R}$, $\max_x f(x)$

or $\min_x f(x)$

Line search methods
↳ Newton
↳ Quasi-Newton

Taylor's theorem

(1) Pick a direction
(2) Go a certain distance in that direction

related: Simulate annealing

A random optimizer – still general purpose

(*) Statistics!

EM algorithm for maximum likelihood

minorization/majorization

Monte Carlo EM

TOPIC:

DATE:

FILE UNDER:

PAGE:

densities $p(x) = C f(x)$

integral

Course outline

Integration "How to compute an integral"

(*) Integration

~~Analytic approximation~~ — Laplace approx.

Quadrature

(*) Monte Carlo (integration)

(EM algorithm = "avoiding integrals")

Random numbers, rejection sampling

Importance sampling

(*) Markov chain Monte Carlo (MCMC)

draw samples from a posterior distribution

Metropolis-Hastings

Gibbs sampling

Variants / ~~tricks~~ tricks

Miscellaneous

(*) Smoothing

↳ splines, Kernel smoothing, p-splines

linear smoothers

(gams)

(*) Bootstrap

TOPIC: | DATE:

FILE UNDER: | PAGE:

Solving nonlinear equations

$f(x) = 0$ for $x \in [a, b]$

Bisection method.

If $\text{sign}(f(a)) \neq \text{sign}(f(b))$

$\Rightarrow$ Intermediate value thm

Let $f(a) < \gamma < f(b)$. $\exists\ c \in [a,b]$ s.t. $f(c) = \gamma$.

$\Rightarrow$ i.o.w. If f is cont. on $[a,b]$, $f^{-1}([a,b])$ is closed

1. Let $c = \frac{a+b}{2}$
2. If $f(c) = 0$, stop
3. Else if $\text{sign}(f(a)) \neq \text{sign}(f(c))$, $b \leftarrow c$.
   else if $\text{sign}(f(b)) \neq \text{sign}(f(c))$, $a \leftarrow c$.
4. goto 1

For n iterations, size of interval is $2^{-n}(b-a)$

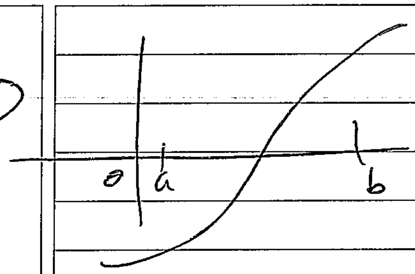
TOPIC:

DATE:

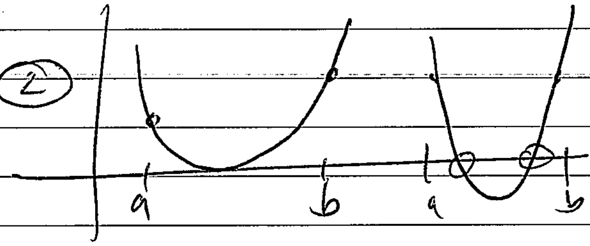
FILE UNDER:

PAGE:

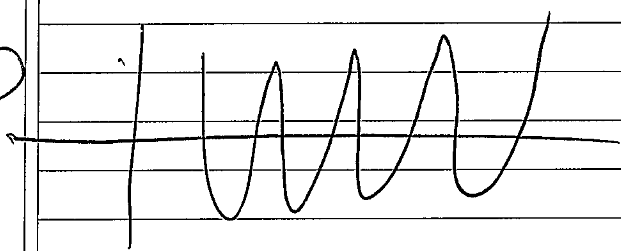
1



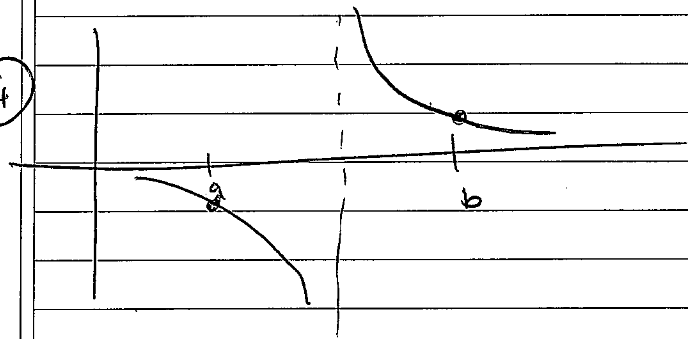
2



3



4



Converge when $|b-a| < \varepsilon$ or

$|f(b) - f(a)| < \varepsilon$. Depends on situation.

Ex.

$\ell(\theta)$ = likelihood

$\ell'(\theta) = 0 \Rightarrow \hat{\theta}$ is MLE

Often want $|\ell(b) - \ell(a)| < \varepsilon$ even if ℓ is flat.

TOPIC: | DATE:

FILE UNDER: | PAGE:

Ex. Quantiles

given cdf $F(x)$, ~~want to find x s.t.~~

and prob $p \in (0,1)$, find x s.t. $F(x) = p$.

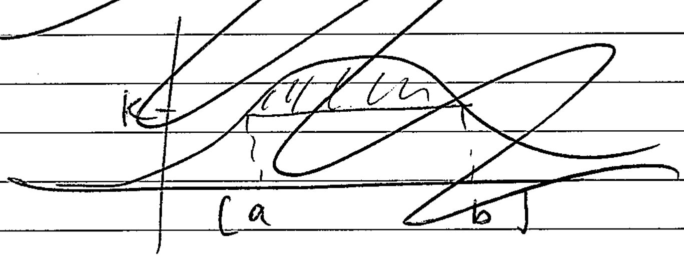
Let $g(x) = F(x) - p$.

Solve $g(x) = 0$

Ex. ~~likelihood intervals~~ ~~HPD intervals~~ Bayesian Credible intervals

Given K,

$S_K \triangleq \{\theta : f(\theta | y) \geq K\}$



Ex. likelihood intervals

Let $f(\theta) = L(\theta) / L(\hat{\theta})$

find $LI = \{\theta : f(\theta) \geq 1/8\}$

Solve $f(\theta) - 1/8 = 0$

TOPIC: Biostat 778

DATE: 2005-11-01

FILE UNDER:

PAGE:

Ex. Bayesian credible interval

Let $S_K = \{\theta : f(\theta | y) \ge K\}$

Bayesian credible interval of level α finds K s.t.

$$\mathbb{P}(\theta \in S_K | y) = \alpha$$

$$\mu([a, b]_K) = \alpha$$

$\mu([a,b]_K) - \alpha = 0$ Solve for K

$[a,b]_K$ Solve for a, b.

For $f : \mathbb{R}^K \to \mathbb{R}$, $K = 2, 3, \ldots, N$

Initial "box" area $= \prod_{i=1}^{K} (b_i - a_i)$

~~At iteration n, area $= \prod_{i=1}^{K} \frac{1}{2}(b_i - a_i)$~~

~~area $= \frac{1}{2^n} \prod_{i=1}^{K} (b_i - a_i)$~~

~~Bisection algorithm: its interval length $\propto \frac{1}{2^n}$~~

At iteration 1, area $= \prod_{i=1}^{K} \frac{1}{2}(b_i - a_i) = \frac{1}{2^K} \prod_{i=1}^{K} (b_i - a_i)$

2 area $= \prod_{i=1}^{K} \frac{1}{2}\frac{1}{2}(b_i - a_i) = \frac{1}{2^{2K}} \prod (b_i - a_i)$

$\vdots$

n area $= \frac{1}{2^{nK}} \prod_{i=1}^{K} (b_i - a_i)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Rates of convergence

Suppose $x_n \to x_\infty^*$ in $\mathbb{R}^k$.

1 Say the convergence is $\boxed{\text{Q-linear}}$ ("linear") if $\exists\, r \in (0,1)$

$$\frac{\|x_{n+1} - x_\infty^*\|}{\|x_n - x_\infty\|} \le r \text{ for all } n \text{ sufficiently large.}$$

Ex. $x_n = 1 + 2^{-n}$ is Q-linear.

$x_\infty = 1$

2 ~~Q-Quadratic~~

Q-superlinear if

$$\lim_{n\to\infty} \frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|} = 0$$

Ex. $x_n = 1 + n^{-n}$ is Q-superlinear

3 Q-Quadratic if

$$\frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|^2} \le M \text{ for all } n \text{ suff. large}$$

Ex. $x_n = 1 + 2^{-2^n}$

TOPIC: Biostat 778

DATE: 2007-10-25

FILE UNDER:

PAGE:

~~Test~~

- test vectors
- homework/reading
- office hours
- no TA

(1) $x_n = 4 + 2^{-n}$

$$\frac{\|x_{n+1}\|}{\|x_n\|} = \frac{4+2^{-(n+1)}}{4+2^{-n}} = \frac{4}{4+2^{-n}} + \frac{2^{-(n+1)}}{4+2^{-n}}$$

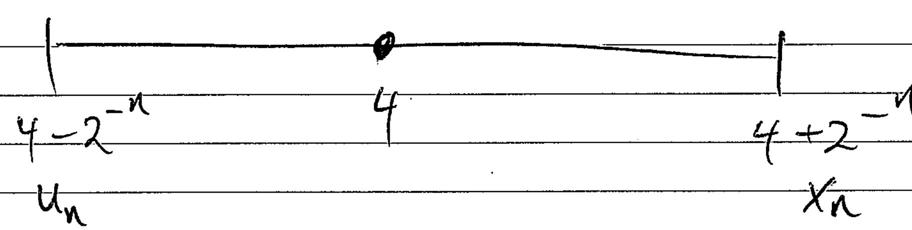
< 1

(2) $x_n = 10000 + 2^{-n}$

$$\frac{x_{n+1}}{x_n} = \frac{1000}{1000 + 2^{-n}} + \frac{2^{-(n+1)}}{1000 + 2^{-n}}$$

$\approx 1 + \varepsilon - 0$

Want to know ratio of errors, not ratio of sequence elements



$$\frac{4-2^{-(n+1)}}{4-2^{-n}}$$

u_n and $x_n \to 4$ at same rate, but

$\frac{x_{n+1}}{x_n} < 1$ and $\frac{|u_{n+1}|}{|u_n|} > 1$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Ex. in bisection algorithm:

Let $X_n = |b_n - a_n|$, i.e. size of interval at iteration n. Then

$$\frac{|X_{n+1} - X_\infty|}{|X_n - X_\infty|} = \frac{X_{n+1}}{X_n} = \frac{2^{-(n+1)}(\cancel{b_0 - a_0})}{2^{-n}(\cancel{b_0 - a_0})}$$

$$= 2^{-n-1+n} = \frac{1}{2} \leq r \in (0,1)$$

Bisection achieves linear convergence

~~Quasi-New~~

Newton's Method — quadratic
Quasi-Newton — superlinear
steepest descent — linear
~~Bisection~~

Functional Iteration

We want to solve $f(x) = 0$ for $f: \mathbb{R}^k \to \mathbb{R}$ and $x \in S \subset \mathbb{R}^k$

Any root of f is a fixed point of $g(x) = f(x) + x$. (There are other functions)

$g(x) = x(f(x)+1)$, $x \neq 0$

Solutions to $f(x)=0$ are fixed points of other functions.

TOPIC:

DATE:

FILE UNDER:

PAGE:

When does functional iteration work?

Sometimes we can take a function f and create a sequence $X_n = f(X_{n-1})$. Depending on f, we can have $X_n \to X_\infty$ where $f(X_\infty) = X_\infty$ (i.e. a fixed point).

<< Shrinking lemma >>

Newton's Method

Solve $f(x) = 0$. ~~Call~~ solution x_∞ and let X_n be our current estimate. By MVT,

$f(X_n) = f'(z)(X_n - X_\infty)$ where

z is b/w X_n and X_∞.

$$\Rightarrow \quad X_\infty = X_n - \frac{f(X_n)}{f'(z)}$$

Since X_∞ and z unknown, ~~let~~ do

$$X_{n+1} = X_n - \frac{f(X_n)}{f'(X_n)}$$

$\Downarrow$

Newton update

<< Proof of Newton's method >>

TOPIC:

DATE:

FILE UNDER:

PAGE:

$S = 1 + \frac{1}{r} + \frac{1}{r^2} + \cdots$

$rS = r + 1 + \frac{1}{r} + \frac{1}{r^2} + \cdots$

$rS = r + S$

$S(r-1) = r$

$S = \frac{r}{r-1} = \frac{1}{1-\frac{1}{r}}$

$\sum_{k=0}^{\infty} \frac{1}{r^k} = 1 + \frac{1}{r} + \frac{1}{r^2} + \cdots$

$a = 1 + \frac{1}{r} + \frac{1}{r^2} + \cdots$

$a = r\left(\frac{1}{r} + \frac{1}{r^2} + \frac{1}{r^3} + \cdots\right)$

Shrinking Lemma

Let M be a closed subset of a c.n.v.s. Let $f: M \to M$ be a map, and assume $\exists\, K$, $0 < K < 1$ s.t. $\forall\, x, y \in M$, we have

$$|f(x) - f(y)| \le K|x - y|.$$

Then f has a unique fixed point. There is a unique pt. $x_0 \in M$ s.t. $f(x_0) = x_0$.

$\Rightarrow$ If $x \in M$, the sequence $\{f^n(x)\}$ is a Cauchy sequence which converges to x_0

Proof:

Given $x \in M$, we have

$$|f^2(x) - f(x)| = |f(f(x)) - f(x)| \le K|f(x) - x|.$$

By induction:

$$|f^{n+1}(x) - f^n(x)| \le K|f^n(x) - f^{n-1}(x)| \le K^n|f(x) - x|$$

And the set of elements $\{f^n(x)\}$ is bounded because

$$|f^n(x) - x| \le |f^n(x) - f^{n-1}(x)| + |f^{n-1}(x) - f^{n-2}(x)| + \cdots + |f(x) - x|$$

$$\le \underbrace{(K^{n-1} + K^{n-2} + \cdots + K)}_{\text{geometric series}}|f(x) - x|$$

$$\le \frac{1}{1-K}|f(x) - x|$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$|f^{m+k}(x) - f^m(x)| \le$

f^m

By induction, Given $m \ge 1$, $K \ge 1$, we have

$$|f^{m+k}(x) - f^m(x)| \le K^m \underbrace{|f^k(x) - x|}_{\le \frac{1}{1-K}|f(x)-x|}$$

$\Rightarrow \exists N$ s.t. if $m, n \ge N$. (say $n = m+k$),

$$|f^{m+k}(x) - f^m(x)| < \varepsilon$$

because $K^m \to 0$ as $m \to \infty$.

$\Rightarrow \{f^n(x)\}$ is a Cauchy sequence. Let x_0 be its limit

Let N be s.t. $\forall n \ge N$, $|f^n(x) - x_0| < \varepsilon$.

Then $|f(x_0) - f^{n+1}(x)| \le K|x_0 - f^n(x)| < \varepsilon$

$\Rightarrow \{f^n(x)\} \Rightarrow f(x_0)$, $\{f^n(x)\} \to x_0$

$\Rightarrow$ $f(x_0) = x_0$, a fixed point

Let x_1 be another fixed point. Then

$|x_1 - x_0| = |f(x_1) - f(x_0)| \le K|x_1 - x_0|$.

Since $0 < K < 1$, $x_1 = x_0$, hence Unique $\square$

TOPIC:

DATE:

FILE UNDER:

PAGE:

Thm:

Let $f \in C^2$ and suppose $\exists\, x_\infty$ s.t. $f(x_\infty) = 0$ and $f'(x_\infty) \neq 0$. Then $\exists\, \delta$ s.t. for any $x_0 \in [x_\infty - \delta, x_\infty + \delta]$, the sequence

$$x_n = g(x_{n-1}) = x_{n-1} - \frac{f(x_{n-1})}{f'(x_{n-1})}$$

converges to x_∞.

Proof:

Note that

$$g'(x) = 1 - \frac{f'(x)f'(x) - f(x)f''(x)}{[f'(x)]^2}$$

$$= \frac{f(x)f''(x)}{[f'(x)]^2}$$

$\Rightarrow g'(x_\infty) = 0$

Since $f \in C^2$, g' is continuous. Therefore

Given $K<1$, $\exists\, \delta > 0$, s.t. $\forall\, x \in [x_\infty - \delta, x_\infty + \delta] = A$

$|g'(x)| < K$.

Also, Given $a, b \in A$,

$$|g(a) - g(b)| \leq |g'(c)||a-b|$$

$$\leq K|a-b| \quad (0<K<1)$$

$\Rightarrow g$ is a shrinking map on A.

$\Rightarrow \exists$ unique x_∞ s.t. $g(x_\infty) = x_\infty$

TOPIC: Biostat 778

DATE: 2005-11-03

FILE UNDER:

PAGE:

Max f = classical / frequentist

$\int f d\mu$ = Bayesian

Convergence rates for shrinking maps.

~~Suppose $g: \mathbb{R}^k \to \mathbb{R}$ and~~

Suppose g satisfies

$|g(x)-g(y)| \leq K|x-y|$

for some $K \in (0,1)$ and any $x, y \in I$, a closed interval.

Also, assume $0 < |g'(x_\infty)| < 1$, where x_∞ is the fixed point. Then $x_n \to x_\infty$ at a linear rate.

Pf: $$\frac{|x_{n+1}-x_\infty|}{|x_n - x_\infty|} = \frac{|g(x_n)-g(x_\infty)|}{|x_n - x_\infty|}$$

Taking limits

$$\lim_{n\to\infty} \frac{|g(x_n)-g(x_\infty)|}{|x_n - x_\infty|} = \underbrace{|g'(x_\infty)|}_{\text{constant} \in (0,1)} > 0$$

$\Rightarrow$ linear convergence

TOPIC:

DATE:

FILE UNDER:

PAGE:

What about Newton's method?

~~Ex~~ Suppose $f \in C^2$, and $\exists\, x_\infty$ s.t. $f(x_\infty) = 0$.

By Taylor's theorem: for some small ε,

(1) $f(x_\infty + \varepsilon) = f(x_\infty) + \varepsilon f'(x_\infty) + \frac{\varepsilon^2}{2} f''(x_\infty) + O(\varepsilon^2)$

$= 0 + \varepsilon f'(x_\infty) + \frac{\varepsilon^2}{2} f''(x_\infty) + O(\varepsilon^2)$

(2) $f'(x_\infty + \varepsilon) = f'(x_\infty) + \varepsilon f''(x_\infty) + O(\varepsilon)$

Newton's method generates the sequence

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

$$\Rightarrow x_{n+1} - x_\infty = x_n - x_\infty - f(x_n)/f'(x_n)$$

Let $\varepsilon_{n+1} = x_{n+1} - x_\infty$, $\varepsilon_n = x_n - x_\infty$.

$$\Rightarrow \varepsilon_{n+1} = \varepsilon_n - f(x_n)/f'(x_n)$$

By +/−, $\varepsilon_{n+1} = \varepsilon_n - \dfrac{f(x_\infty + \varepsilon_n)}{f'(x_\infty + \varepsilon_n)}$

$$\Rightarrow \varepsilon_{n+1} \doteq \varepsilon_n - \frac{\varepsilon_n f'(x_\infty) + \varepsilon_n^2 f''(x_\infty)/2}{f'(x_\infty) + \varepsilon_n f''(x_\infty)}$$

$$= \frac{\cancel{\varepsilon_n f'} + \varepsilon_n^2 f'' - \cancel{\varepsilon_n f'} - \varepsilon_n^2 f''/2}{f' + \varepsilon_n f''}$$

$$= \varepsilon_n^2 \left(\frac{f''/2}{f' + \varepsilon_n f''} \right)$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$$\frac{\varepsilon_{n+1}}{\varepsilon_n^2} \approx \frac{f''(x_\infty)/2}{f'(x_\infty) + \varepsilon_n f''(x_\infty)}$$

$$\approx \frac{f''(x_\infty)/2}{f'(x_\infty)} \qquad \varepsilon_n \downarrow 0$$

$\Rightarrow \exists$ some $M < \infty$ s.t.

$$\left| \frac{\varepsilon_{n+1}}{\varepsilon_n^2} \right| \leq M, \quad \forall n \text{ suff. large.}$$

$\Rightarrow$ Quadratic convergence.

Of course we need $f''(x_\infty)$ exists and $f'(x_\infty) \neq 0$.

In practice, we ignore assumptions / conditions.
Use Newton's method as a "black box". Caveat Emptor.

Pro: Very fast in neighborhood of truth
Direct multivariate generalization

Con: Need to evaluate f'
Can be unstable.

TOPIC:

DATE:

FILE UNDER:

PAGE:

We want $\hat{\theta}$, the value of θ that maximizes $\ell(\theta)$. Assume that $\hat{\theta}$ is the unique root of $\ell'(\theta)$. Solve $\ell'(\theta) = 0$ (likelihood equations)

<u>Newton's method</u>

$$\underset{K\times 1}{\theta_{n+1}} = \theta_n - \underbrace{[\ell''(\theta_n)]}_{K\times K}^{-1} \underset{K\times 1}{\ell'(\theta_n)}$$

$\Rightarrow$ May be easier/better to solve

$$\underbrace{[\ell''(\hat{\theta}_n)]}_{A}\underset{x}{\theta_{n+1}} = \underbrace{[\ell''(\theta_n)]\,\theta_n - \ell'(\theta_n)}_{b}$$

Than try to invert $\ell''(\theta_n)$.

At convergence we have, in addition to mle $\hat{\theta}$,

$\ell'(\hat{\theta})$: score statistic

$-\ell''(\hat{\theta})$: observed information

The obs. information is related to the covariance matrix of limiting normal dist. of $\hat{\theta}$, i.e.

$$\sqrt{n}\left([-\ell''(\hat{\theta})^{-1/2}](\hat{\theta} - \theta_0)\right) \longrightarrow N(0, I)$$

for $n \longrightarrow \infty$.

TOPIC:

DATE:

FILE UNDER: Binary logistic regression

PAGE:

$\log p_i - \log 1-p_i = x_i^T\beta$

$\log p_i - x_i^T\beta$

$Y_i \sim Bernoulli(p_i)$, $i = 1, \ldots, n$

$$logit(p_i) = \log \frac{p_i}{1-p_i} = x_i^T\beta$$

$$p_i = \frac{e^{x_i^T\beta}}{1+e^{x_i^T\beta}}$$

$$L(\beta) \propto \prod_{i=1}^{n} p_i^{Y_i}(1-p_i)^{1-Y_i}$$

$$= \exp\left(\sum_{i=1}^{n} Y_i \log p_i + (1-Y_i)\log(1-p_i)\right)$$

$$= \exp\left(\sum Y_i\left(x^T\beta - \log(1+e^{x_i^T\beta})\right) + (1-Y_i)\left(-\log(1+e^{x_i^T\beta})\right)\right)$$

$$\frac{\partial \ell}{\partial \beta} = \sum Y_i\left[x_i - \frac{x_i e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right] + (1-Y_i)\left(\frac{-x_i e^{x_i^T\beta}}{1+e^{x_i^T\beta}}\right)$$

$$= \sum Y_i\left[x_i - x_i p_i\right] + (1-Y_i)\left(-x_i p_i\right)$$

$$(Y_i - 1)(x_i p_i)$$

$$= \sum Y_i\left[x_i - x_i p + x_i p_i\right] - x_i p_i$$

$$= \sum Y_i x_i - x_i p_i$$

$$= \sum x_i(Y_i - p_i) = X^T(Y-p)$$

$\beta_{new} = \beta_n + \left[-X^T W X\right]^{-1} X^T(Y-p)$

$\ell''(\beta) = -X^T W X$

$W = diag\left[p_i(1-p_i)\right]$

TOPIC:

DATE:

FILE UNDER:

PAGE:

$$\ell'(\beta) = X^T(Y - p)$$

$$\ell''(\beta) = -X^T W X$$

$$W \hookrightarrow \mathrm{diag}[p_i(1-p_i)]$$

$$\beta_{n+1} = \beta_n - [-X^T W_n X]^{-1} [X^T(Y - p_n)]$$

Newton update

TOPIC:

DATE:

FILE UNDER:

PAGE:

For exp. families w/canonical link, $l''(\theta) = \mathbb{E}\, l''(\theta)$ so Newton and Fisher scoring are same

General Purpose Minimization

Given a function $f: \mathbb{R}^k \to \mathbb{R}$, we want to find $\min_{x \in S} f(x)$ where $S \subset \mathbb{R}^k$.

usually $\iff f'(x) = 0$

Line search methods

Given f and a current estimate of the location of the minimum x_n, we want to

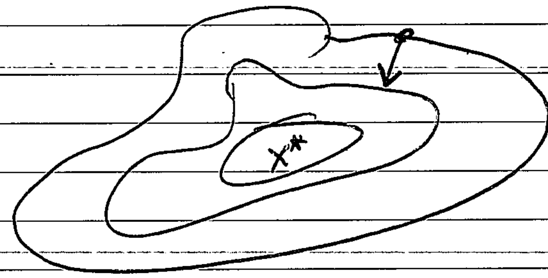
1. Choose a direction p_n (vector)
2. Solve $\min_{\alpha > 0} f(x_n + \alpha p_n)$
   - don't need exact min. Rather compute some candidates and choose the best one.
3. $x_{n+1} = x_n + \alpha p_n$

Choosing Direction

Most obvious is steepest-descent : ~~$f'(x_n)$~~

$-f'(x_n)$: direction along which f decreases most rapidly

→ orthogonal to contours of f



TOPIC:

DATE:

FILE UNDER:

PAGE:

Newton direction:

By Taylor's Theorem:

$$f(x_n+p) \approx \underbrace{f(x_n) + p^T f'(x_n) + \tfrac{1}{2} p^T f''(x_n)\, p}_{m_n(p)}$$

minimize over p $m_n(p)$ $\Rightarrow$ $p_n = \left[-f''(x_n)\right]^{-1} f'(x_n)$

{Newton direction has "natural" step length of 1 but that can be modified}

$\Rightarrow$ $x_{n+1} = x_n + \left[-f''(x_n)^{-1}\right] f'(x_n)$

That's familiar!

Similarly, Quasi-Newton

$$f'(x_n) - f'(x_{n-1}) = B_n(x_n - x_{n-1})$$

$$p_n = B_n^{-1} f'(x_n)$$

$\downarrow$

B_n satisfies a "secant condition"

Coordinate descent

If f is K-dimensional, we minimize along individual coordinate dimensions in a cyclic fashion.

$\Rightarrow$ method of alternating variables

$\Rightarrow$ cyclic coordinate descent

$\Rightarrow$ "deterministic Gibbs sampling"

$\Rightarrow$ backfitting

TOPIC:

DATE:

FILE UNDER:

PAGE:

Variations of Newton's method

Again, solve $\ell'(\theta) = 0$

Newton's method: $\theta_{n+1} = \theta_n - \ell''(\theta_n)^{-1}\,\ell'(\theta_n)$

In general: $\theta_{n+1} = \theta_n - B_n^{-1}\,\ell'(\theta_n)$

(Fisher) scoring: replace ℓ'', the observed information with the expected information matrix. (sometimes easier to compute)

Used to fit GLMs where it is equiv. to IRLS (with canonical link fcn)

Quasi-Newton: Replace ℓ'' with a "secant-like" approximation

(*) $\underbrace{\ell'(\theta_n) - \ell'(\theta_{n-1})}_{y_n} = B_n\,\underbrace{(\theta_n - \theta_{n-1})}_{s_n}$

Solve (*) for B_n (not unique, many ways).

unlike $-\ell''(\hat{\theta})$, B_n is not a valid estimate of $Var(\hat{\theta})$!!

Popular method due to Broyden, Fletcher, Goldfarb, and Shanno (BFGS). Also DFP (Davidon, Fletcher, Powell)

$\Rightarrow$ In 1-D case, there is a unique solution

~~$A s_n = y_n$~~ ~~"secant equation"~~

TOPIC: | DATE:

FILE UNDER: | PAGE:

$$\underbrace{\ell'(\theta_n) - \ell'(\theta_{n-1})}_{Y_n} = B_n \underbrace{(\theta_n - \theta_{n-1})}_{S_n}$$

$B_n S_n = Y_n$ "secant equation"

~~DFP?~~ $\to$ infinite solutions

$\cancel{B_n = (I - \gamma_n Y Y}$

Add'l constraint: Find B closest to previous one, and symmetric.

$$B_n = \arg\min_B \| B - B_{n-1} \| \text{ subj. to.}$$

$$B = B^T \text{ and } B s_n = Y_n$$

$\Rightarrow$ solution is DFP method

Let $H_n = B_n^{-1}$

~~Let~~ Solve $\min_H \| H - H_{n-1} \|$

subj. to $H = H^T$ and $H y_n = S_n$

$\Rightarrow$ solution is BFGS method

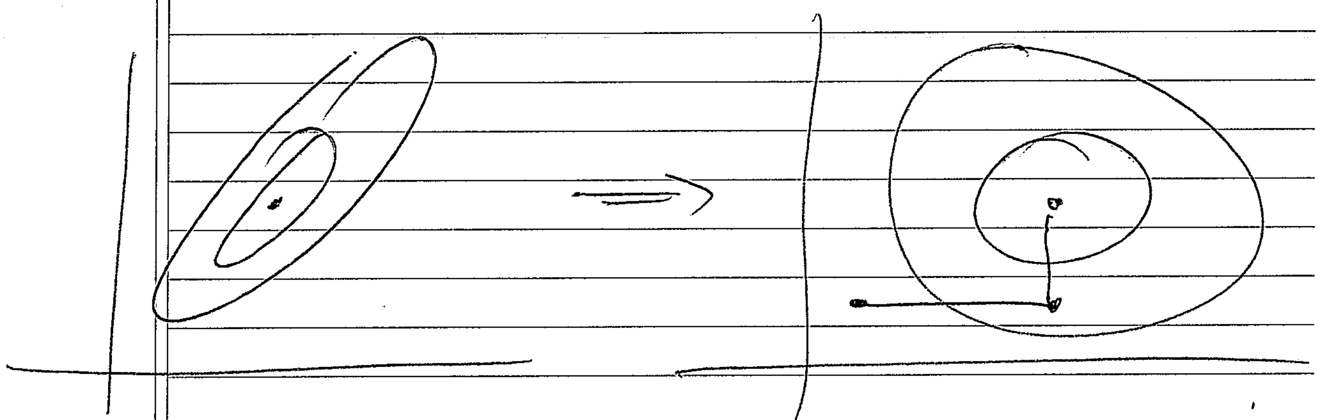
TOPIC:

DATE:

FILE UNDER:

PAGE:

Conjugate Gradient



Evalute $f_0 = f(x_0)$, $f'_0 = f'(x_0)$

Let $p_0 = -f'(x_0)$

1. find $\min_{\alpha > 0} f(X_n + \alpha p_n) \Rightarrow \alpha_n$

   Set $X_{n+1} = X_n + \alpha_n p_n$

2. Eval $f'(X_{n+1}) = f'_{n+1}$

3. let

$$\beta_{n+1} = \frac{f'^{T}_{n+1} f'_{n+1}}{f'^{T}_{n} f'_{n}}$$

[Fletcher — Reeves]

4. $p_{k+1} = -f'_{n+1} + \beta_{n+1} p_n$

TOPIC:

DATE:

FILE UNDER:

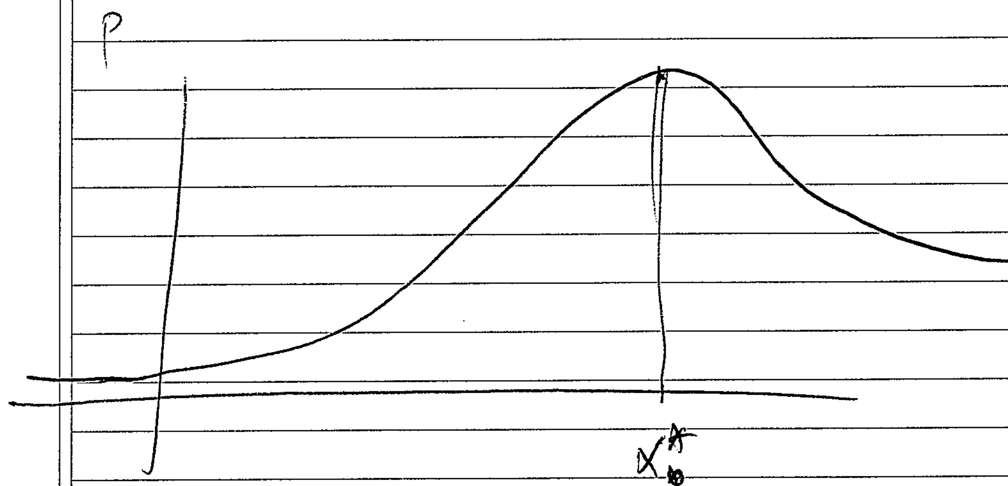
PAGE:

Polak-Ribière:

$$\beta_{k+1} = \frac{f'^{T}_{n+1}\,(f'_{n+1} - f'_{n})}{f'^{T}_{n}\, f_{n}}$$

$$p_0 = -f'(x_0) = -f'_0$$

$$p_1 = -f'_1 + \frac{f'^{T}_1 f'_1}{f'^{T}_0 f'_0}\left(-f'_0\right)$$



$$f(x) = x^2 \cancel{+ cy^2 + cxy}$$

$$\cancel{f'(x) = 2x} \qquad f' = \begin{pmatrix} 2x+y \\ 2y+x \end{pmatrix}$$

$$p_0 = -2x_0 \Rightarrow x_1 = x_0 + \alpha(-2x_0)$$

$$p_1 = -2x_1 + \frac{4x_1^2}{4x_0^2}\left(-2x_0\right)$$

$$= -2x_1 + \frac{x_1^2}{x_0^2}\left(-2x_0\right)$$

TOPIC:

DATE:

FILE UNDER:

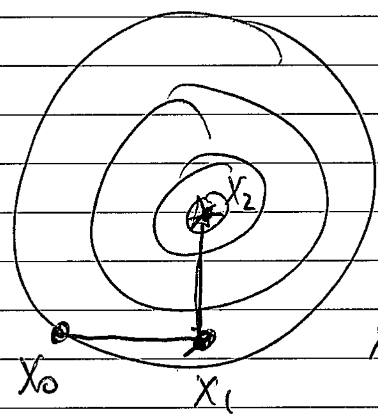
PAGE:

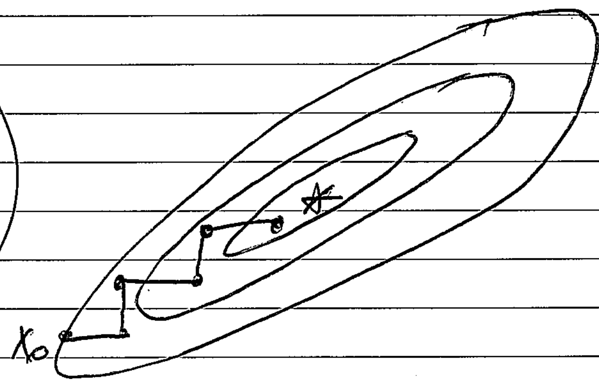
Coord. descent

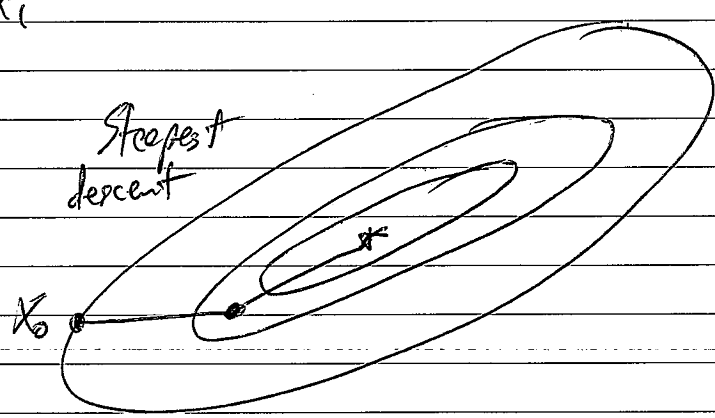
Coordinate descent can be very slow (slower than steepest descent).

But:

1. Does not require calc. of f' or f''
2. Often easier to do many 1-D mins than one K-D min.
3. Convergence can be good if coordinates are loosely coupled







TOPIC:

DATE:

FILE UNDER:

PAGE:

Step-length selection

Given a step direction p_n, how far to go?

Let $\phi(\alpha) = f(x_n + \alpha p_n)$.

Find $\min_{\alpha > 0} \phi(\alpha)$

$\Rightarrow$ Too hard!

Roughly speaking:

(1) Choose initial α_0. If

(*) $\phi(\alpha_0) \le \phi(0) + c\,\alpha_0\,\phi'(0)$ — sufficient decrease condition

then stop. $(c \approx 10^{-4})$

(2) Otherwise, make quadratic approximation to $\phi(\alpha)$ — called ϕ_q — and let α_1 be the minimizer of ϕ_q.

If α_1 satisfies (*), stop.

(3) Otherwise, make cubic approximation to ϕ, called ϕ_c, and let α_2 minimize ϕ_c.

If α_2 satisfies (*), stop

else repeat (3).

Cubic functions are good for approximating fcns with much curvature.

optimize() in R uses polynomial (cubic) approximation.

TOPIC:

DATE:

FILE UNDER:

PAGE:

~~Simu~~ Simulated annealing → more later

Simplex Method

It's clear that in minimizing f, there is a tradeoff b/w Knowledge of f and rate of convergence

~~f^0 only~~ ~~partial~~

Knowledge	Method ~~#~~		
f only	CCD (simulated annealing, simplex)	sublinear (?) (linear)	Slow
f' only	steepest descent	linear	
partial f''	Quasi-Newton, Fisher Scoring	super linear	↓
Full f''	Newton	quadratic	fast

~~The EM algorithm~~

Fisher Scoring — Poisson regression

$Y_i \sim \text{Poisson}(\mu_i)$ } $Y \sim \text{Poisson}(\exp(X^T\beta))$

$\eta_i = \log \mu_i = X_i^T\beta$

$\eta = X^T\beta$

$V(\mu) = \mu$

$\mu = \exp(X^T\beta)$

$\eta = \log \mu$

$\frac{d\eta}{d\mu} = \frac{1}{\mu}$

① Start with $\hat{\mu}_0$

② Set $\hat{z} = \hat{\eta} + (Y - \hat{\mu}) / \hat{\mu}$

(adjusted response, working response)

$\hat{\eta} = \log \hat{\mu}$