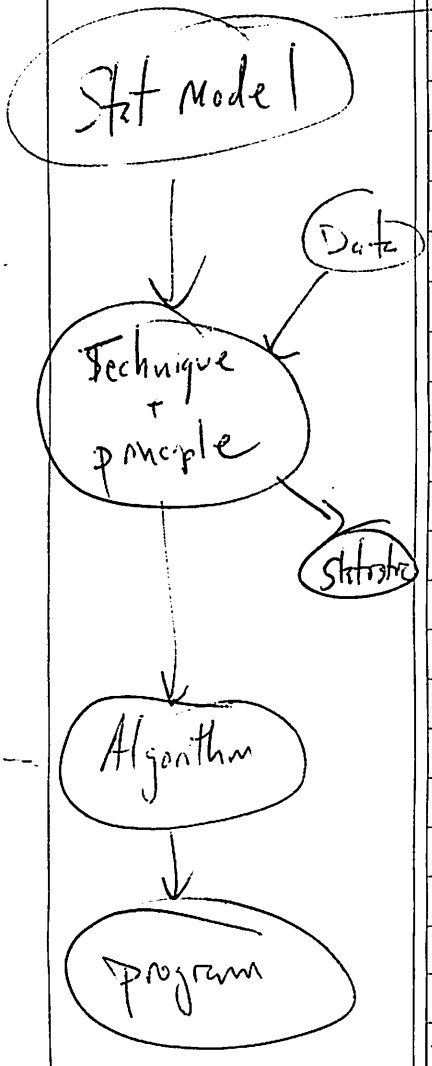
TOPIC: Biostat 778

DATE: 2005-10-27

FILE UNDER:

PAGE:



(1)

Solutions to nonlinear equations

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

- Bisection
- functional iteration
- Newton's method

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 f = min $-f$

So don't worry about it.

This course is about

1. maximizing a function
2. integrating a function

Course outline

(*) Solving nonlinear equations (root finding)

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

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

$K = 2, 3, \ldots, N$

Maximizing

(*) 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

~~Deterministic Algorithms~~

General idea:

Entire Class/course

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*

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

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

$f(x_{n-1}) = f'(c)(x_{n-1} - x_{n-2})$

~~Course outline~~

Course outline

Integration
"How to compute an integral"

Miscellaneous

derivative: $p(x) = C f(x)$ ← 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

(*) Smoothing

↳ splines, Kernel smoothing, p-splines
linear smoothers
(gams)

(*) Bootstrap

$\frac{1}{2\pi}$ $\frac{1}{2\pi\sigma}\exp\left(-\frac{1}{2\sigma}x^2\right)$

Computation

① Linear regression

$$\hat{\beta} = (X'X)^{-1}X'y$$

R uses QR decomp

$$X = QR$$

$$\hat{\beta}\,\cancel{(X'X)} = (\cancel{R'R)^{-1}R'}Q'y$$

②

$$X'X\beta = X'y$$

$$R'Q'QR\beta = R'Q'y$$

$$R'R\beta = R'Q'y$$

$$R\beta = Q'y$$

② Multivariate normal

$$\varphi(x|\mu,\Sigma)$$

$$= \underbrace{-\frac{1}{2}\log|\Sigma|}_{\det} - \frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)$$

$$\varphi(z|\Sigma) = -\frac{p}{2}\log 2\pi - \log|\Sigma| - \frac{1}{2}z'\Sigma^{-1}z$$

$\Sigma = R'R$ (cholesky)

$z'\Sigma^{-1}z = z'(R'R)^{-1}z$: $z'R^{-1}R'^{-1}z$

$z'LL'z$

$= z'R'^{-1}R^{-1}z$

x' x

$= (R'^{-1}z)'R^{-1}z$

$R'x = z$ backsolve

$x = R'^{-1}z$

$R^{-1}RR^{-1}R^{-1}$

$\Rightarrow z'\Sigma^{-1}z = x'x$

$z'(R'R)^{-1}z$

$z'R^{-1}R'^{-1}z$

$= (R^{-1}z)'R^{-1}z$

~~$x = R'^{-1}$~~

$x = R'^{-1}z$

$x' = z'R$

Solving non linear equations

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

Bisection method.

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

$\Rightarrow$ 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 $\text{sgn}(f(a)) \neq \text{sgn}(f(c))$, $b \leftarrow c$.
else if $\text{sgn}(f(b)) \neq \text{sgn}(f(c))$, $a \leftarrow c$.

(4) goto (1)

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

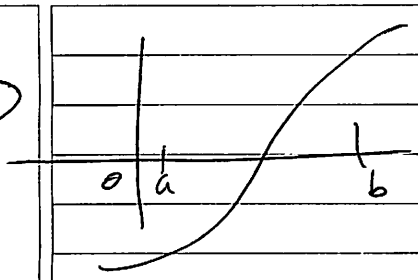
TOPIC:

DATE:

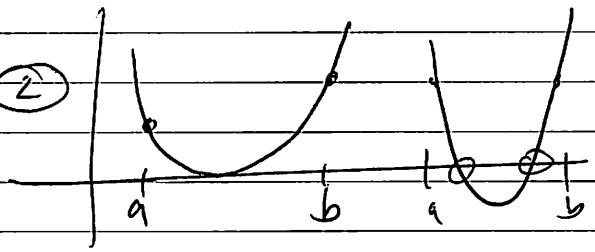
FILE UNDER:

PAGE:

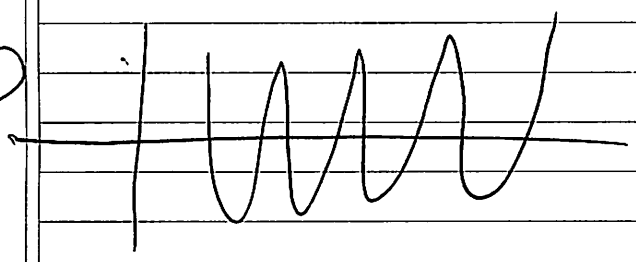
(1)



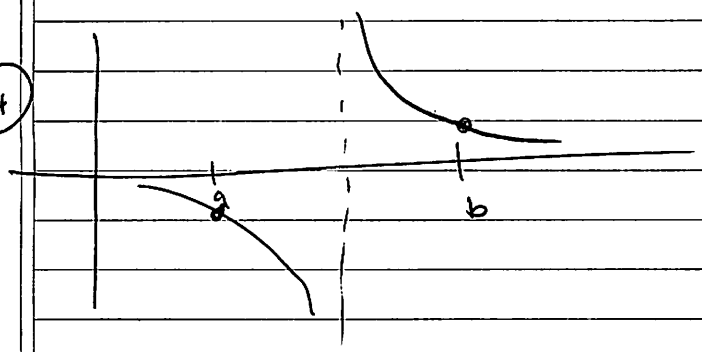
(2)



(3)



(4)



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

Ex.

$l(\theta)$ = likelihood

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

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

Ex. Quantiles

Given cdf $F(x)$, ~~want to find x s.t.~~ and prob $p \in (0,1)$, find x s.t. $F(x) = p$.

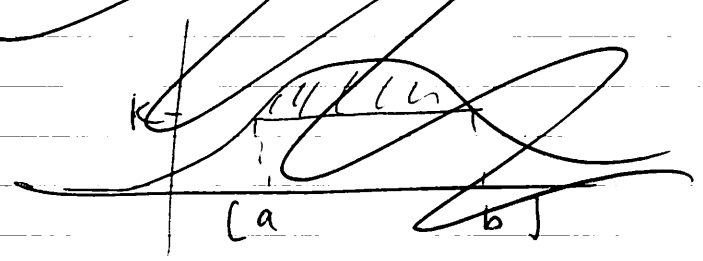
Let $g(x) = F(x) - p$.

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

Ex. Bayesian credible interval

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

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

$$R(\theta \in S_K | y) = \alpha$$

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

$\mu([a,b]_K) - \alpha = 0$ 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 $= \frac{1}{2^{nK}} \prod_{i=1}^{K} (b_i - a_i)$

Bisection algorithm: ~~& interval length $\alpha \frac{1}{2^n}$~~

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

Rates of convergence

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

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

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

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

$x_\infty = 1$

(2) 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

$$\frac{\|x_{n+1} - x_\infty\|}{\|x_n - x_\infty\|^2} \leq M \quad \text{for all } n \text{ sufficiently large}$$

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

Ex. In bisection algorithm:

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

$$\frac{|x_{n+1} - x_\infty|}{|x_n - x_\infty|} = \frac{x_{n+1}}{x_n} = \frac{2^{-(n+1)}(b_0 - a_0)}{2^{-n}(b_0 - a_0)}$$

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

Bisection achieves linear convergence

~~Quasi-Newton~~

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:

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).

When does fixed point iteration work?

<< 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, 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:

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

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

$rS = r + S$

$S(r-1) = S$

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

$S(r-1) = r$

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

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

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

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

Pf:

Given $x \in M$, we have

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

By induction:

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

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

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

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

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

$\frac{1}{1-K}$

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

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

$a = r(\cdots)$

TOPIC:

DATE:

FILE UNDER:

PAGE:

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

f^m

By induction, Given $m \geq 1$, $k \geq 1$, we have

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

$\Rightarrow \exists N$ s.t. if $m, n \geq N$ (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. $\forall n \geq N$, $|f^n(x) - x_0| < \varepsilon$.

Then

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

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

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

Let x_1 be another fixed point. Then

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

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_∞.

Pf:

Note that

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

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

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

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. $g(x_\infty) = x_\infty$

Max f = classical / frequentist

$\int f d\mu$ = Bayesian

No real assumptions of g here abt g

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

Taking limits

$$\lim_{n \to \infty} \frac{|g(x_n) - g(x_\infty)|}{|x_n - x_\infty|} = |g'(x_\infty)| > 0$$

constant $\in (0,1)$

$\Rightarrow$ linear convergence

What about Newton's method?

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

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

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

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

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

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

$\Rightarrow \exists$ 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.

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)

Newton's method

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

$K \times 1$

$\Rightarrow$ May be easier/better to solve

$$\underbrace{\left[\ell''(\theta_n)\right]}_{A}\theta_{n+1} = \underbrace{\left[\ell''(\theta_n)\right]\theta_n - \ell'(\theta_n)}_{b}$$

$A \quad x \quad = \quad b$

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 θ, i.e.

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

for $n \to \infty$.

TOPIC DATE

FILE UNDER PAGE

Binary logistic regression

$\log p_i - \log(1-p_i) = x_i \beta$

$\log i - x_i \beta$

$y_i \sim \text{Bernoulli}_i(p_i)$, $i = 1, \dots, n$

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

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

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

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

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

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

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

$$(y_i - 1)(x_i p_i)$$

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

$$= \sum y_i x_i - x_i p_i$$

$$= \sum x_i (y_i - p_i) = X^T (y - p)$$

$$= \sum x_i p_i (1-p_i) x_i$$

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

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

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

TOPIC:

DATE:

FILE UNDER:

PAGE:

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

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

$$\hookrightarrow \operatorname{diag}\left[p_i(1-p_i)\right]$$

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

Newton update

For exp. families w/canonical link, $\ell''(\theta) = \mathbb{E}\,\ell''(\theta)$ 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

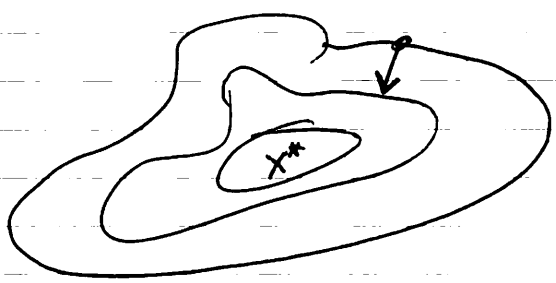
1. Choose a direction p_n (vector)
2. Solve $\min_{\alpha > 0} f(x_n + \alpha p_n)$
   - ↳ don't need exact min. Rather compute some candidates and choose the best one.
3. $x_{n+1} = x_n + \alpha p_n$

Choosing Direction

Most obvious ⇒ steepest-descent: ~~$\nabla f(x_n)$~~

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

↳ orthogonal to contours of f



Newton direction:

By Taylor's Theorem:

$$f(x_n + p) \approx f(x_n) + p^T f'(x_n) + \frac{1}{2} p^T f''(x_n) p$$

$$\Downarrow$$

$$m_n(p)$$

Minimize $m_n(p)$ over 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}

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

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.

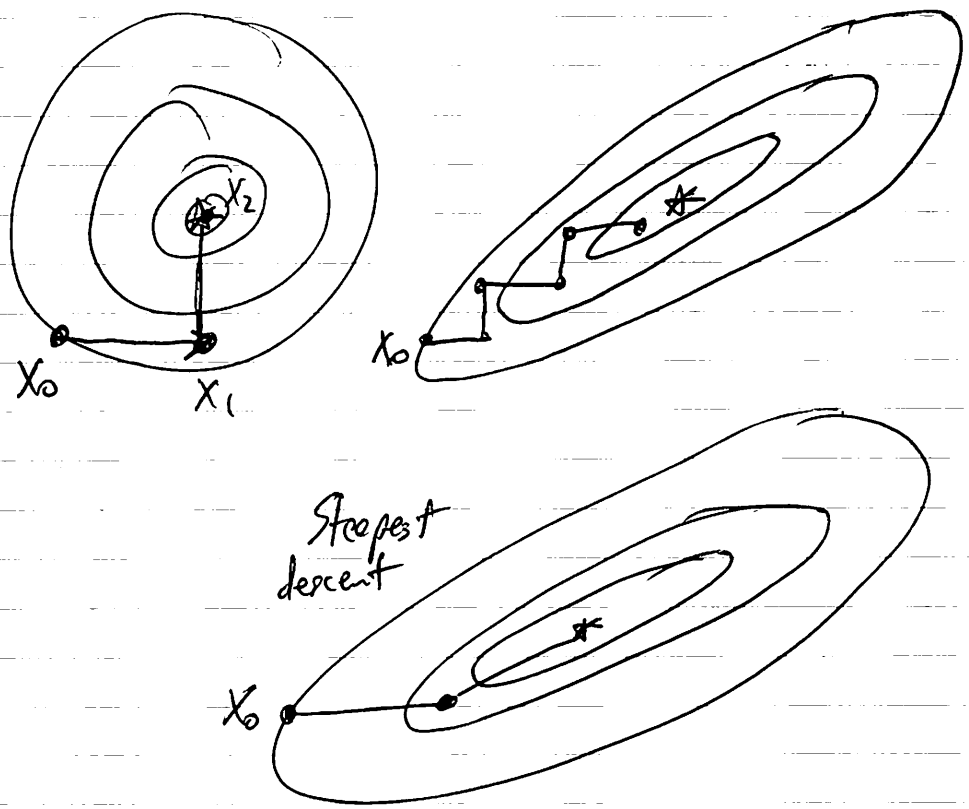
$\Rightarrow$ Method of alternating variables
$\Rightarrow$ cyclic coordinate descent
$\Rightarrow$ "deterministic Gibbs sampling"
$\Rightarrow$ backfitting

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



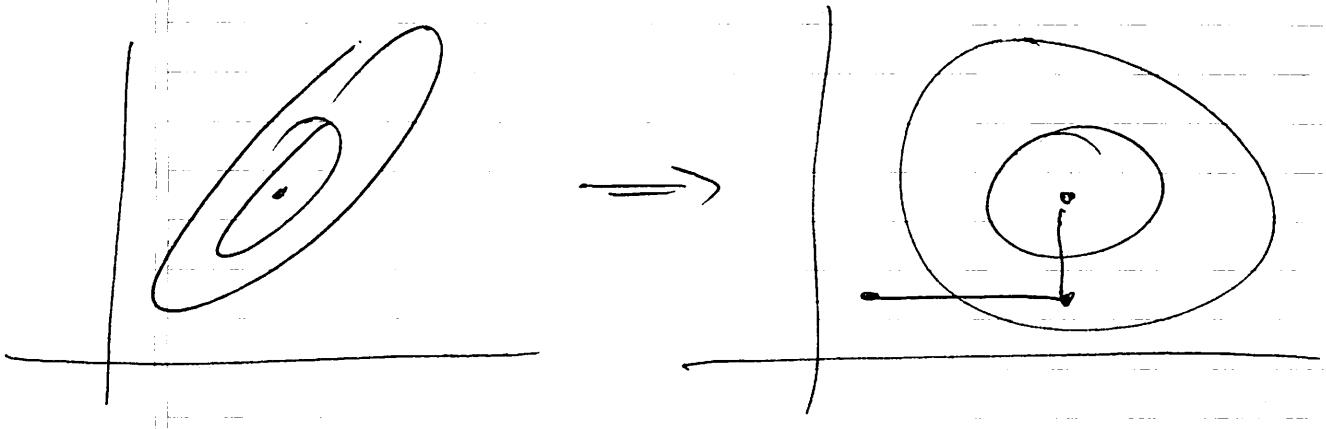
TOPIC

DATE

FILE UNDER:

PAGE

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 $\beta_{n+1} = \dfrac{f'^{T}_{n+1} f'_{n+1}}{f'^{T}_{n} f'_{n}}$ [Fletcher — Reeves]

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

TOPIC

DATE

FILE UNDER

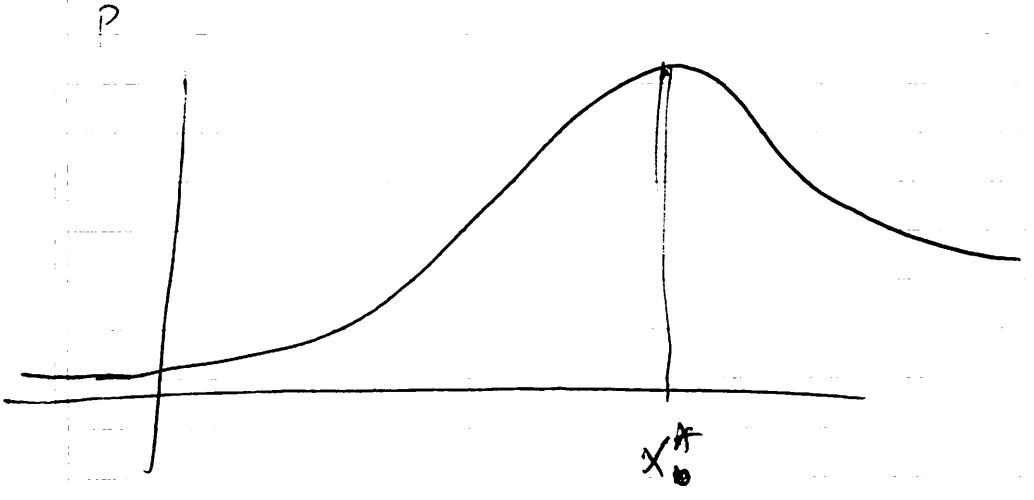
PAGE

Polak-Ribiere:

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

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

$$p_1 = -f'_1 + \frac{f'^{T}_1 f'_1}{f'^{T}_0 f_0}(-f'_0)$$



$f(x) = x^2$

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

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

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

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

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'' w. a "secant-like" approximation

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

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

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

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

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

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

$B_n s_n = y_n$ "secant equation"

DFP: → infinite solutions

$B_n = (I - \ldots)$

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

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

$$B = B^T \quad \text{and} \quad B s_n = y_n$$

⇒ solution is DFP method

Let $H_n = B_n^{-1}$

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

$$\text{subj. to} \quad H = H^T \quad \text{and} \quad H y_n = s_n$$

⇒ solution is BFGS method

$f'(x)$

$\phi(\alpha) = f(x)$

$\phi'(0) = [f(x+\alpha p)]' = [f'(x+\alpha p)]^T$

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

(*) $\phi(\alpha_0) \leq \phi(0) + c_1 \alpha_0 \phi'(0)$ — 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 functions with much curvature.

optimize() in R uses polynomial (cubic) approximation

NAME

FILE UNDER

PAGE

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

Knowledge	Method	
f only	CCD (simulated annealing, simplex)	sublinear (?) (linear)
f' only	steepest descent	linear
partial f''	Quasi-Newton Fisher Scoring	super linear
FULL f''	Newton	quadratic

Slow ↓ 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 $\hat{\mu}$.

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

(adjusted response, working response)

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

Fisher Scoring for GLMs

$Y \sim$ [illegible] , $p(y \mid \mu)$ $\quad E[y] = \mu$

$g(\mu) = X\beta$ $\quad Var(y) = V(\mu)$

$$g(y) \doteq g(\mu) + (y-\mu) g'(\mu)$$

$\downarrow$

z (working response)

$$\eta = g(\mu)$$

(1) Start with $\hat{\mu}_i$

(2) Set $z_i = \eta_i + (y_i - \hat{\mu}_i)\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$

(3) Weighted regression of z on X, i.e. solve

$$X^T W X \beta_n = X^T W z$$

$$\beta_n = (X^T W X)^{-1} X^T W z$$

where $$W = \left[\left(\frac{\partial \eta_i}{\partial \mu_i}\right)^2 V(\mu_i)\right]^{-1}$$

ex: $y_i \sim \text{Poisson}(\mu_i)$

$$g(\mu_i) = \log \mu_i = \eta_i$$

$$\frac{\partial \eta_i}{\partial \mu_i} = \frac{1}{\mu_i} \qquad V(\mu_i) = \mu_i$$

TOPIC:

DATE:

FILE UNDER:

PAGE:

(1) set $\hat{\mu}_i$

(2) $z_i = \log \mu_i + (Y_i - \mu_i)/\mu_i$

(3) Regress z on X where

$$W = \left[\frac{1}{\hat{\mu}_i^2}\hat{\mu}_i\right]^{-1} = \hat{\mu}_i$$

Poisson regression via Newton

$$L(\beta) \propto e^{-\mu} \mu^Y$$

$$\ell(\beta) = Y \log \mu - \mu = Y X\beta - \exp(X\beta)$$

$$\ell'(\beta) = X^T Y - X^T \exp(X\beta) \qquad X^T Y - X^T \mu = X^T(Y-\mu)$$

$$\ell''(\beta) = -X^T \underbrace{\exp(X\beta)}_{\mu} X \qquad -X^T W X$$

$z = \eta + (Y-\mu)/\mu$

$\mu z = \mu\eta + Y - \mu$

$Y - \mu = \mu z - \mu\eta$

$$\beta_{n+1} = \beta_n + \left(-X^T W X\right)^{-1}\left(-X^T(Y-\mu)\right)$$

$$= \beta_n + (X^T W X)^{-1} X^T (\mu z - \mu\eta)$$

$$= \beta_n + (X^T W_n X)^{-1} X^T W_n (z_n - \eta_n)$$

$$= (X^T W X)^{-1} X^T W z + \underbrace{\left(\beta_n - (X^T W X)^{-1} X^T W \eta\right)}$$

$$= (X^T W X)^{-1} X^T W z \qquad = 0 \qquad = \log \mu = X\beta$$

$\eta = g(\mu) = \log \mu$

$\eta = z - \frac{Y-\mu}{\mu}$

$X\beta$

$\mu = \exp$

$l_{1r} =$

The EM algorithm

EM stands for Expectation-Maximization.

Originally outlined in DLR (1977) but ideas go much further back. DLR unified many different ideas and gave examples of the broad applicability of EM.

EM is not an "algorithm". It is an algorithm for creating other algorithms.

Generalized Additive Models (Hastie + Tibshirani)

Usual linear model has:

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

GAMS say:

$$Y_i = \alpha + S_1(X_{1i}) + S_2(X_{2i}) + \cdots + S_p(X_{pi}) + \varepsilon_i$$

$$\Rightarrow Y = \alpha + \sum_{j=1}^{p} S_j(X_j) + \varepsilon \qquad (*)$$

where $S_j(\,)$ are smooth.

$S_j(\,)$ can be any kind of "smoother", even a mixture of different kinds

ex. Smoothers: parametric splines, smoothing splines, penalized splines, loess, running lines, running median

GAM algorithm: → $Y = (Y_1, \to Y_n)$

Given model $Y = \alpha + \sum_{j=1}^{p} s_j(X_j) + \varepsilon$

(1) Initialize $\alpha = \frac{1}{n}\sum_{i=1}^{n} y_i$, $s_j = s_j^0 = ?\ 0$

(2) For $j = 1, 2, \to p$

Let $r_j = Y - \alpha - \sum_{k \neq j} s_k(X_k)$

~~where $s_k = (s_k(X_{k1}), s_k(X_{k2}), \ldots, s_j(X_{jn}))$~~

where $s_k = (s_k(X_{k1}), s_k(X_{k2}), \to s_k(X_{kn}))$

i.e. $s_k()$ evaluated at data points for X_k

$s_j = \text{Smooth}(r_j \mid X_j)$

$\hat{s}_j = s_j - \frac{1}{n}\sum_{i=1}^{n} s_j(x_{ij})$

(3) Evaluate $\Delta = \sum_{j=1}^{p} \| s_j - s_j^0 \|$

or $\Delta = \dfrac{\sum_{j=1}^{p} \| s_j - s_j^0 \|}{\sum \| s_j^0 \|}$

If $\Delta < \varepsilon$, stop.

Else set $s_j^0 = s_j$ for $j = 1, \to p$.

Goto (2)

Local Scoring

Set adjusted response

$$z_i = \eta_i + (Y_i - \mu_i)\frac{\partial \eta_i}{\partial \mu_i}$$

Fit an additive model for

$z \sim x_1 \rightarrow x_p$ with observation weights

$$W_i = \left[\frac{\partial \eta_i}{\partial \mu_i}\right]^{-2} V(\mu_i)^{-1}$$

GAM algorithm = "backfitting"

~ Alternating conditional expectation
~ Cyclic coordinate descent

⇒ linear convergence algorithm

⇒ sidesteps curse of dimensionality by additivity constraint (big!)

Boosting

Outcomes are $Y_i \in \{-1, 1\}$ with covariates X_i.

Given $g(x) \in \mathbb{R}$, then let

$$f(x) = \text{sign}(g(x)) \in \{-1, 1\}$$

e.g. logistic regression

$$g(x) = \text{logit}\ \frac{\mu(x)}{1-\mu(x)} = x'\beta \in (-\infty, \infty)$$

$$\Rightarrow \quad f(x) = \text{sign}(x'\beta)$$

Margin: $Y_i f(X_i)$

$Y_i f(X_i) > 0$ good!

$Y_i f(X_i) < 0$ bad!

Loss functions: Classifiers minimize some loss function

Binomial deviance:

$$L(y, f(x)) = \log\left(1 + e^{-2yf(x)}\right)$$

Squared error:

$$L(y, f(x)) = (y - f(x))^2$$

Exponential:

$$L(y, f(x)) = e^{-yf(x)}$$

Ada Boost

(1) Set $w_i = \frac{1}{n}$ for $i = 1 \to n$

(2) For $m = 1, \to M$

(a) Fit classifier $f_m(x)$ to training data weighted by w_i

(b) Compute

$$err_m = \frac{\sum_{i=1}^{n} w_i \mathbb{1}\{y_i \neq f(x_i)\}}{\sum_{i=1}^{n} w_i}$$

(c) Let $c_m = \log \frac{1 - err_m}{err_m}$

(d) Update $w_i = w_i \exp\left(c_m \mathbb{1}\{y_i \neq f(x_i)\}\right)$ for $i = 1, \to n$

(3) Final classifier is $F(x) = \text{sign}\left[\sum_{m=1}^{M} f_m(x)\right]$

Boosting

Friedman, Hastie, Tibshirani, 2000 Annals

(1) AdaBoost

→ why this?

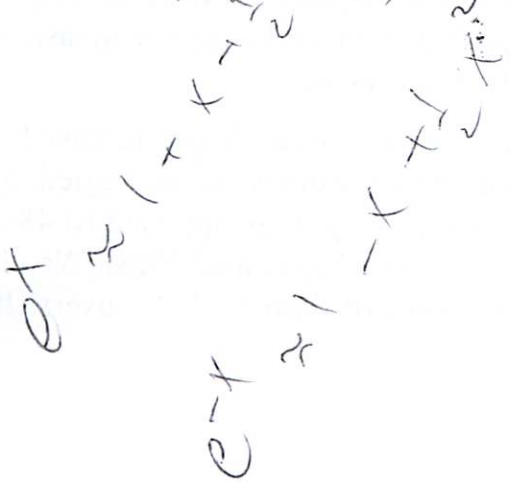
$\mathbb{E}[e^{-yF(x)}]$ is minimized by $\frac{1}{2}\log\frac{\mathbb{P}(y=1|x)}{\mathbb{P}(y=-1|x)}$

Pf: $\frac{\partial}{\partial F}\mathbb{E}[e^{-yF(x)}|x] = -e^{-F(x)}\mathbb{P}(y=1|x) + e^{F(x)}\mathbb{P}(y=-1|x) = 0$

$-e^{-F(x)}\mathbb{P}(y=1|x) = -e^{F(x)}\mathbb{P}(y=-1|x)$

$\frac{\mathbb{P}(y=1|x)}{\mathbb{P}(y=-1|x)} = \frac{-e^{F(x)}}{-e^{-F(x)}} = e^{2F(x)}$

$\frac{1}{2}\log\frac{\mathbb{P}(y=1|x)}{\mathbb{P}(y=-1|x)} = F(x)$



(2) AdaBoost $\Longrightarrow$ Additive logistic regression model via Newton-like steps for minimizing $\mathbb{E}[e^{-yF(x)}]$

Pf: Let $J(F) = \mathbb{E}[e^{-yF(x)}]$. Given x and c, improve $F(x)$ by taking a step $F(x) + c f(x)$.

$\Rightarrow$ Newton's method was $x_{n+1} = x_n + \underbrace{f''(x_n)^{-1} f'(x_n)}_{\text{scalar}} = x_n + (1) f(x_n)$ (function)

$f \in \{1, -1\}$

$$J(F + cf) = \mathbb{E}[e^{-y(F+cf)}] = \mathbb{E}[e^{-yF} e^{-ycf}]$$

$$\approx \mathbb{E}[e^{-yF}(1 - ycf + \tfrac{1}{2}y^2c^2f^2)]$$

$y^2 = 1$, $f^2 = 1$

$$= \mathbb{E}[e^{-yF}(1 - ycf + \tfrac{1}{2}c^2)]$$

We want $f = \underset{f}{\operatorname{argmin}}\, \mathbb{E}[e^{-yF}(1 - ycf + \tfrac{1}{2}c^2)]$

for a given x $= \underset{f}{\operatorname{argmin}}\, \mathbb{E}_w[1 - ycf + \tfrac{1}{2}c^2]$

For $C > 0$,

$$\min \mathbb{E}_w\left[1 - yCf + \tfrac{1}{2}C^2\right]$$

$$\Leftrightarrow$$

$$\max \mathbb{E}_w[yf] = \frac{\mathbb{E}\left[e^{-yF} yf \,|\, x\right]}{\mathbb{E}\left[e^{-yF} \,|\, x\right]}$$

$$\frac{\partial}{\partial f}\mathbb{E}_w[yf] = \mathbb{E}_w[y \,|\, x] = 0$$

$$= f(x)\mathbb{E}_w[y \,|\, x]$$

$$= f(x)\left[1\,\mathbb{P}_w(y=1 \,|\, x) - 1\,\mathbb{P}_w(y=-1 \,|\, x)\right]$$

$$= f(x)\left[\mathbb{P}_w(y=1 \,|\, x) - \mathbb{P}_w(y=-1 \,|\, x)\right]$$

$$f(x) = \begin{cases} 1 & \text{if } \mathbb{P}_w(y=1 \,|\, x) \geq \mathbb{P}_w(y=-1 \,|\, x) \\ -1 & \text{otherwise} \end{cases}$$

$\Rightarrow$ Minimizer

To get c, we minimize L_w for fixed $f(x)$

$$J(F+cf) = \mathbb{E}\left[e^{-YF - cYf}\right]$$

$$= \mathbb{E}_w\left[e^{-Ycf}\right]$$

$$= e^{-cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right] + e^{cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right]$$

$$\frac{\partial}{\partial c} J(F+cf) = -f e^{-cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right] + f e^{cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right] = 0$$

$$e^{cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right] = e^{-cf}\,\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right]$$

$$e^{2cf} = \frac{\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right]}{\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right]}$$

$$cf = \frac{1}{2}\log\frac{\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right]}{\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right]}$$

$$\Rightarrow \quad c = \frac{f}{2}\log\frac{\mathbb{E}_w\left[\mathbb{1}\{Y=1\}\right]}{\mathbb{E}_w\left[\mathbb{1}\{Y=-1\}\right]} = \frac{1}{2}\log\frac{1-\mathrm{err}}{\mathrm{err}}$$

$$\mathrm{err} = \mathbb{E}_w\left[\mathbb{1}\{Y \neq f(x)\}\right]$$

The new $F(x)$ is then

$$F_\alpha(x) = F(x) + c f(x)$$

$$= F(x) + \frac{1}{2} \log \frac{1-err}{err} f(x)$$

new weights are

$$e^{-yF} \Rightarrow e^{-yF} e^{-cf(x)y}$$

because $-yf(x) = 2\,\mathbb{1}\{y \neq f(x)\} - 1$, new weights are

$$e^{-yF}\left(\exp\left(\frac{1}{2}\log\frac{1-err}{err}\left(2\,\mathbb{1}\{y \neq f(x)\} - 1\right)\right)\right)$$

$$= e^{-yF}\left[\exp\left(\log\frac{1-err}{err}\,\mathbb{1}\{y \neq f(x)\}\right)\right](K)$$

$\Rightarrow$ Same as Discrete AdaBoost

Biostat 778: Homework 1

November 16, 2016

1 Improving consistent estimators

Let $\tilde{\theta}_n$ be an estimator of θ such that $\sqrt{n}(\tilde{\theta}_n - \theta) \to \mathcal{N}(0, \Sigma(\theta))$ and $\Sigma(\theta) < \infty$. Let $\hat{\theta}_n$ be the maximum likelihood estimator of θ. Let $\tilde{\theta}_n^{(1)}$ be single iteration of Newton's method applied to $\tilde{\theta}$, i.e.

$$\tilde{\theta}_n^{(1)} = \tilde{\theta}_n - \ell''(\tilde{\theta}_n)^{-1}\ell'(\tilde{\theta}_n)$$

Show that $\tilde{\theta}_n^{(1)}$ is asymptotically equivalent to the MLE.

2 Logistic Regression with Penalization

Write a function that can fit a logistic regression model while allowing for an L^2 penalty on the parameters. That is, if $\ell(\beta)$ is the log-likelihood for the parameter vector β, then you want to maximize the penalized log-likelihood

$$\ell^*(\beta, \lambda) = \ell(\beta) - \lambda\beta'\beta$$

Write your own implementation of Newton's method to optimize the penalized likelihood. The output of the function should be the maximum (penalized) likelihood estimates of β and the asymptotic standard errors for each of the elements of $\hat{\beta}$.

TOPIC

DATE

FILE UNDER

PAGE

Summary of Minimization

$\min_x f(x)$ for $f: \mathbb{R}^k \to \mathbb{R}$, $x \in \mathbb{R}^k$

Line Search:

(1) Steepest descent

$$x_{n+1} = x_n + \alpha \underbrace{[-f'(x_n)]}_{\text{direction of steepest descent}}$$

α: scalar step length

$\Rightarrow$ linear convergence

(2) Newton

$$x_{n+1} = x_n + 1 \underbrace{[f''(x_n)]^{-1}[-f'(x_n)]}_{\text{Newton direction}}$$

Step length $= 1$

$\Rightarrow$ quadratic convergence

$\rightarrow$ in stat, estimate of asymptotic covariance

(3) Quasi-Newton

$$x_{n+1} = x_n + \alpha B_n [-f'(x_n)]$$

(a) $B_n = \arg\min_B \|B - B_{n-1}\|$

$B = B^T$ and $f'(x_n) - f'(x_{n-1}) = B_n(x_n - x_{n-1})$

④ Conjugate Gradient ("Modified steepest descent")

$$x_{n+1} = x_n + \alpha p_n$$

$$\downarrow$$

$$p_{n+1} = -f_n'(x_{n+1}) + \frac{\|f'(x_{n+1})\|^2}{\|f'(x_n)\|^2} p_n$$

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

$$p_1^T p_1 = f_0' f_1 + \frac{\|f_1'\|^2}{\|f_0\|^2} f_1^T f_0$$

Coordinate Descent:

$$x_{n+1}^{(k)} = \underset{x^{(k)}}{\arg\min}\, f(x_n^{(1)}, x_n^{(2)}, \ldots, x_n^{(k)}, \ldots, x_n^{(p)})$$

$\Rightarrow$ has "descent property", non-increasing

Method of steepest descent/ascent uses

$$\theta_{n+1} = \theta_n - \alpha \ell'(\theta_n)$$

Where α is chosen so that θ_{n+1} has a larger likelihood value than θ_n.

Do this ⟹

Knowledge of ℓ''

Summary:

① Newton: Fastest (quadratic convergence), requires calculating ℓ'', gives asymptotic $\mathrm{Var}(\hat{\theta})$, unstable if starting value too far. ✓

② Scoring: Super linear convergence, but equiv. to Newton if ℓ'' does not depend on Y_0 (true in many common cases), requires calculating $E[\ell'']$, can be unstable, but often quite stable in typical statistics apps. ~

③ Quasi-Newton: Superlinear convergence, does not require ℓ'' or $E[\ell'']$, more stable than Newton, No estimate of $\mathrm{Var}(\hat{\theta})$ ~

~~④ Steepest descent: linear convergence, very stable~~

④ Steepest descent: linear convergence, stable X

TOPIC

DATE

FILE UNDER

PAGE

Secant method

If f' is difficult to compute (almost always) or we are lazy (always) the the secant method provides an approximation.

Newton step/Secant method

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

$$\Rightarrow f'(x_n) \approx \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$$

Note: Need 2 starting values.

Pro: easy as pie

Con: Convergence only superlinear.

$x_n - x_{n-1} =$
$x_n - x_\infty + x_\infty - x_{n-1}$
$\varepsilon_n \quad \varepsilon_{n-1}$

$\varepsilon_{n+1} = \varepsilon_n -$
$\frac{f(x_\infty + \varepsilon_n)(x_n - x_{n-1})}{f(x_\infty + \varepsilon_n) - f(x_\infty + \varepsilon_{n-1})}$

$\varepsilon_n -$
$\frac{(\varepsilon_n f' + \frac{\varepsilon_n^2}{2} f'')(\quad)}{\varepsilon_n f' - \frac{\varepsilon_n^2}{2} f'' - \varepsilon_{n-1} f' - \frac{\varepsilon_{n-1}^2}{2} f''}$

$= \varepsilon_n -$
$\frac{(\varepsilon_n f' + \frac{\varepsilon_n^2}{2} f'')(\varepsilon_n - \varepsilon_{n-1})}{(\varepsilon_n - \varepsilon_{n-1}) f' - \left(\frac{\varepsilon_n^2}{2} - \frac{\varepsilon_{n-1}^2}{2}\right) f''}$

Statistics application

$\ell(\theta): \mathbb{R}^k \to \mathbb{R}$, log-likelihood, $\theta = (\theta_1, \ldots, \theta_k)$

$\ell'(\theta): \mathbb{R}^k \to \mathbb{R}^k$, gradient

$\ell''(\theta): \mathbb{R}^k \to \mathbb{R}^{k \times k}$, hessian