

# Convergence Order

How fast does  $|E_t| \rightarrow 0$  ?

Convergence order  $\beta$  if

$$\lim_{t \rightarrow \infty} \frac{|e_{t+1}|}{|e_t|^\beta} = C$$

for some  $C \neq 0$ ,  $\beta > 0$ .

Higher  $\beta \Rightarrow$  more improvement  
at each step.

E.g. if  $e_t = 0.001$   $\beta = 1$   $c = 1/2$

then  $e_{t+1} \approx 0.0005$

$e_{t+2} \approx 0.00025 \dots$

$\approx 2.5$  steps to get to  $1.00001$

Q1  $\beta = 2$ ,  $c = 1$

$e_{t+1} = 0.00000025$

etc.

# Fisher Scoring

$-l''(\underline{\theta})$  not necessarily (pos/neg) - definite.

But  $\underline{I}(\underline{\theta}) = E_{\underline{\theta}}(-l''(\underline{\theta}))$

is at least pos. semi-definite.

So can use  $-\underline{I}(\underline{\theta})$  instead of  $-l''(\underline{\theta})$

Update:  $\underline{\theta}_{t+1} = \underline{\theta}_t + l'(\underline{\theta}_t) \underline{I}(\underline{\theta}_t)^{-1}$

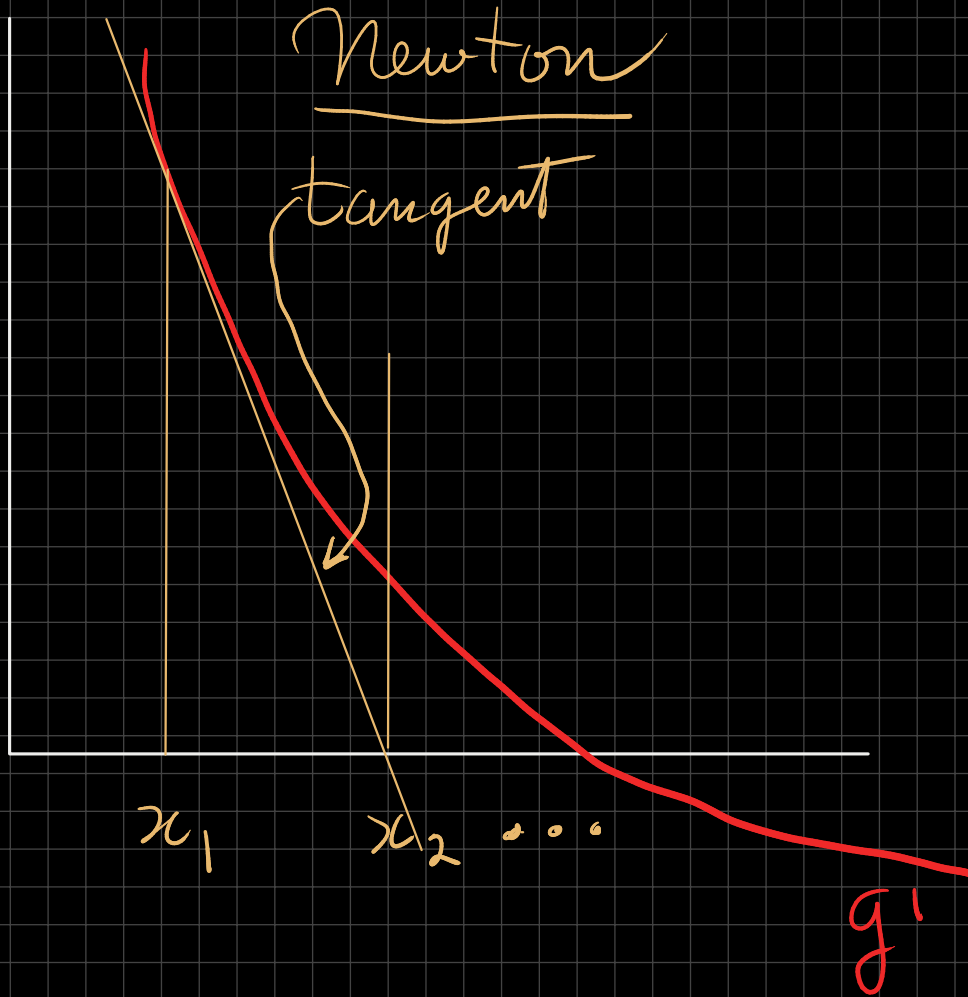
Secant Method: For  $\theta \in \mathbb{R}$

$$\text{Use } \frac{|g'(\theta_t) - g'(\theta_{t-1})|}{\theta_t - \theta_{t-1}}$$

To guess  $l''(\theta_t)$

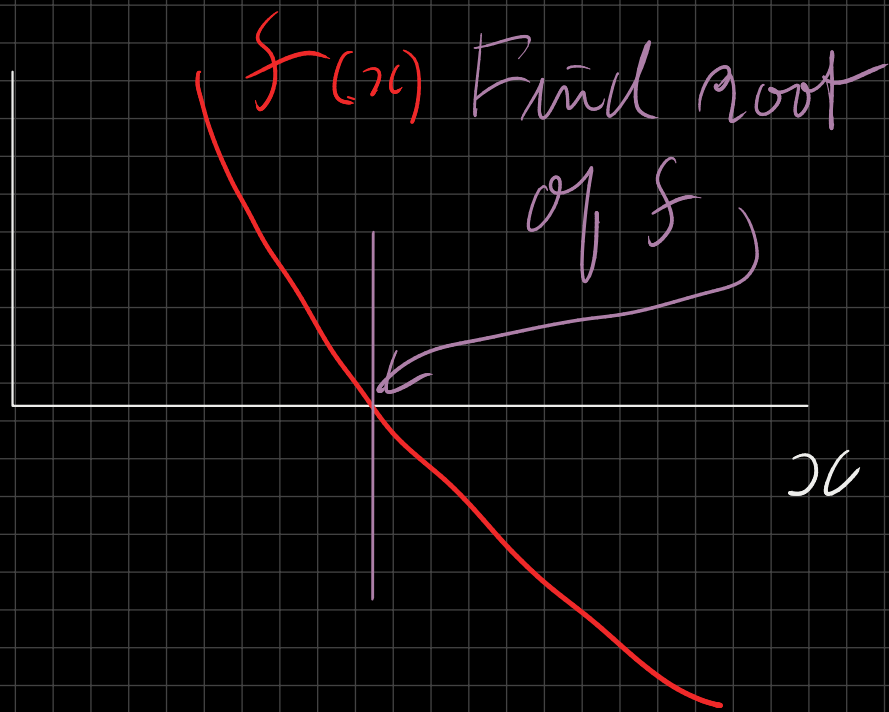
- Need 2 starting values.  $\theta_1, \theta_2$
- Will adapt for  $\mathbb{R}^p$

# Why "secant"

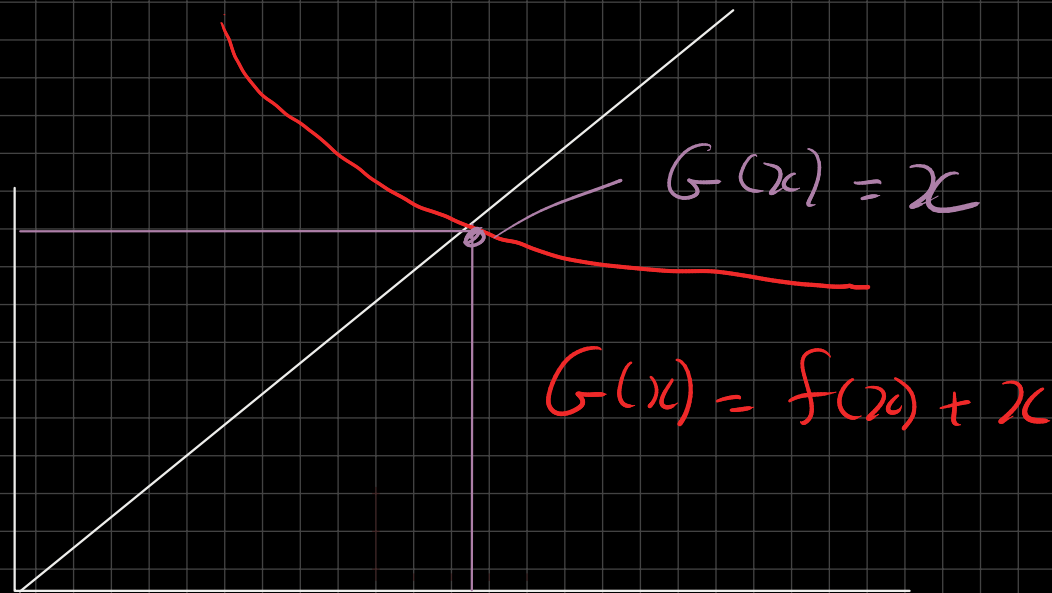




# Fixed point iteration

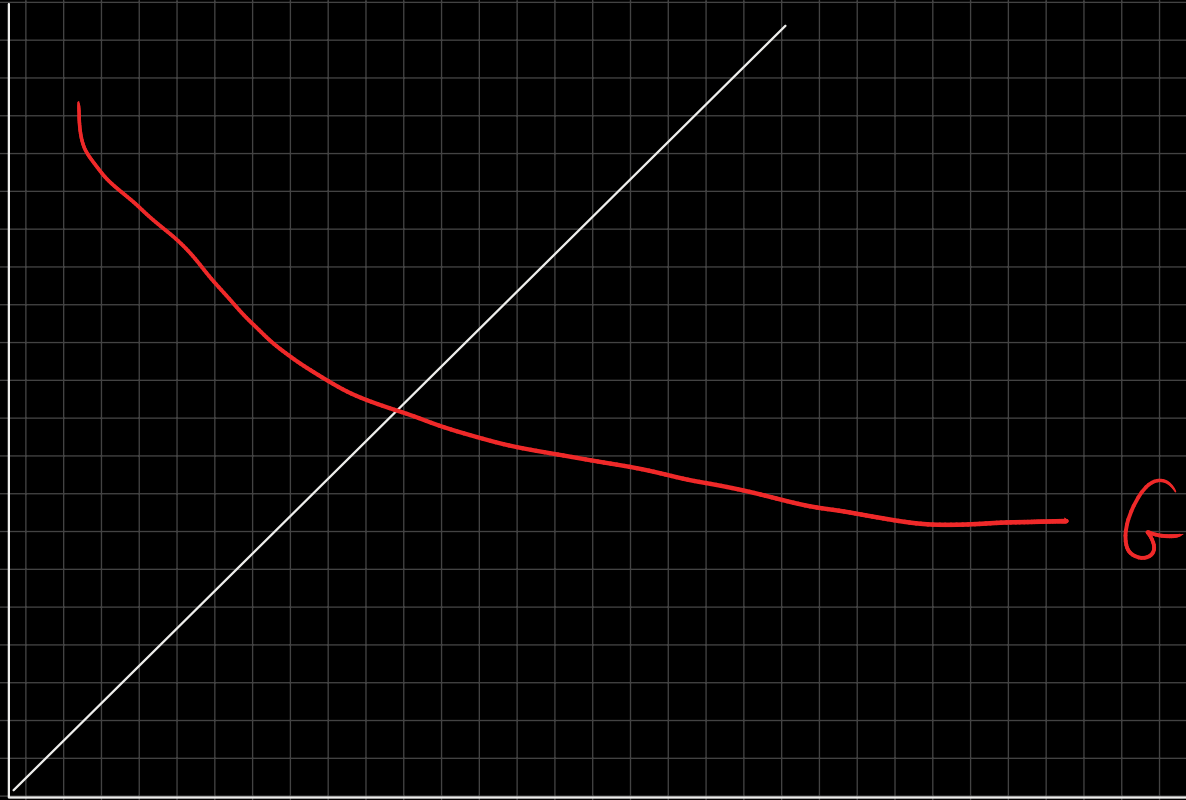


Same as



Idea:  $x_{t+1} = G(x_t)$   
until  $x_{t+1} \approx x_t$

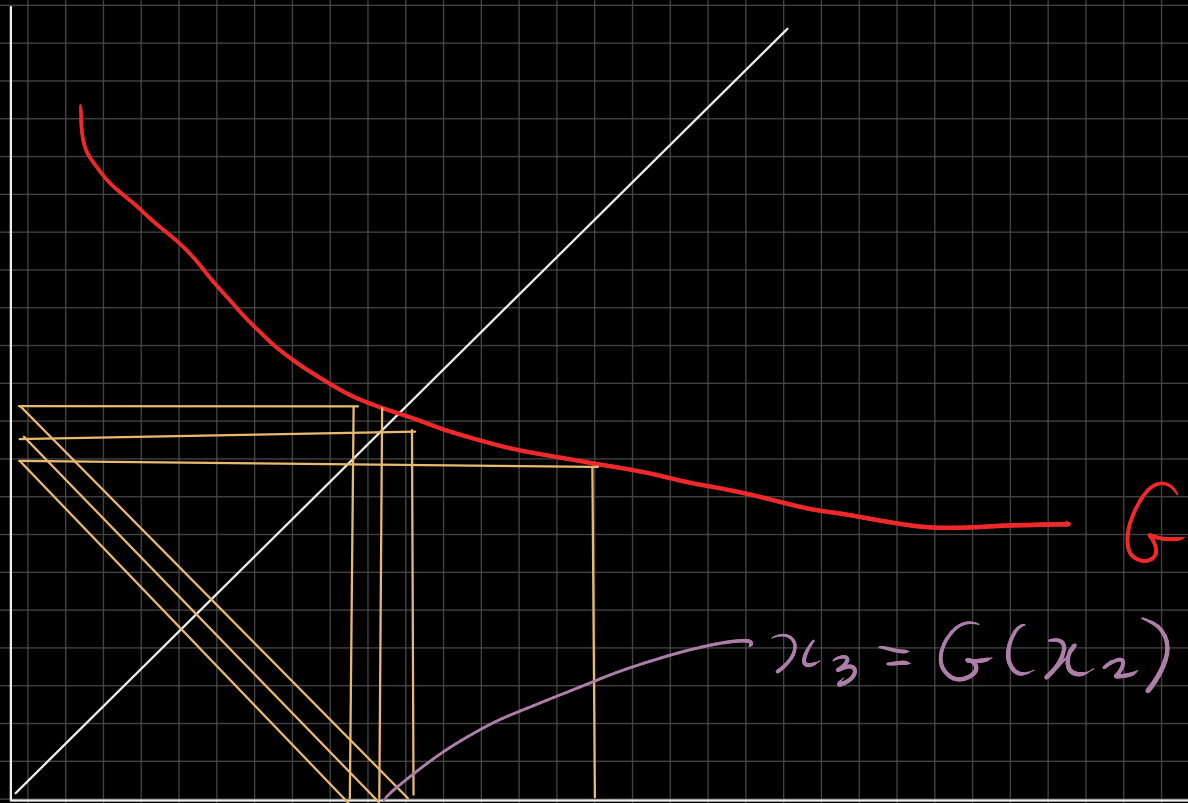
How does this work?



$$G(x) = f(x) + x$$



How does this work?

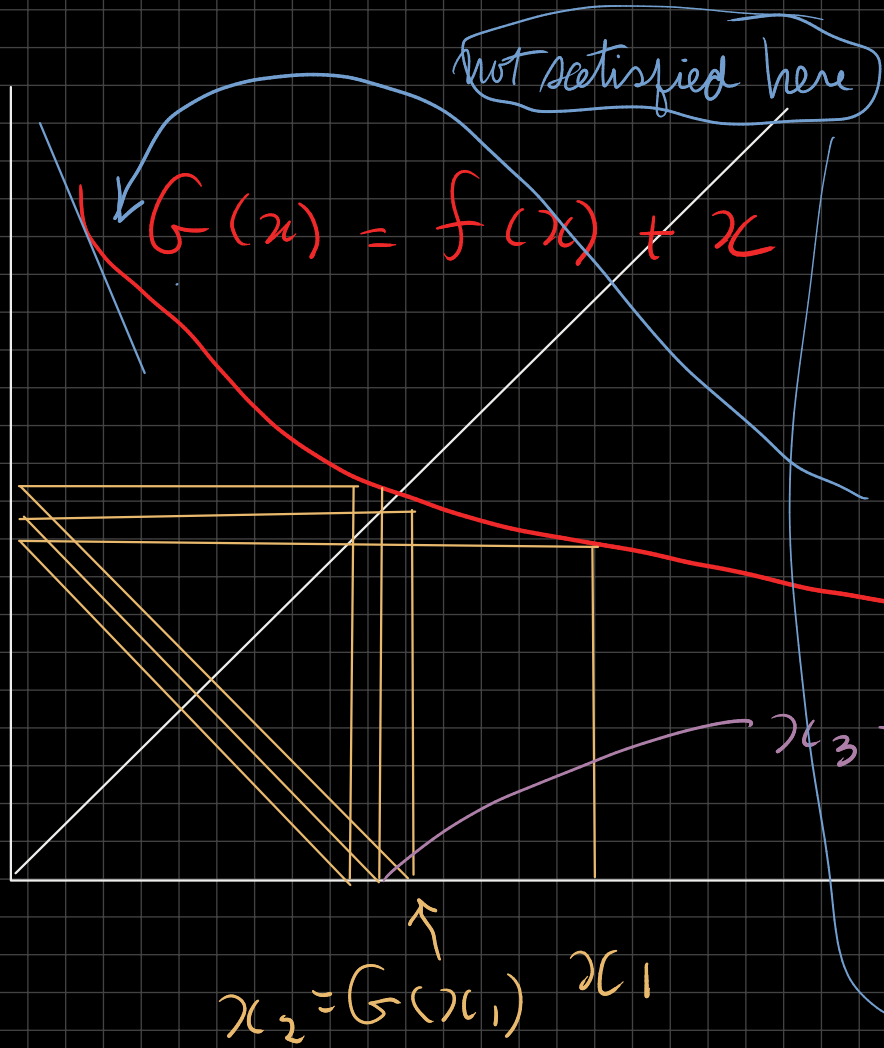


$$G(x) = f(x) + x$$

$$x_3 = G(x_2)$$

$$x_2 = G(x_1)$$

How does this work?



When does this work?

$G$  contractive on  $[a, b]$

1)  $G([a, b]) \subset [a, b]$

Note:  $a = -\infty, b = +\infty$  ok

2) Lipschitz condition

$$\left\{ \begin{array}{l} |G(x_1) - G(x_2)| \leq \lambda |x_1 - x_2| \end{array} \right.$$

on  $[a, b]$  for some

$$\lambda \in [0, 1)$$

so we would have to restrict  $x$  to  $[a, b]$  on which Lipschitz is satisfied

## OR: Scaling

Instead of  $G(x) = f(x) + x$

Could use  $G(x) = \alpha f(x) + x$  for  $\alpha \neq 0$

since  $G(x) = \alpha f(x) + x = x$

$$\Rightarrow \alpha f(x) = 0 \Rightarrow f(x) = 0$$

---

Lipschitz condition is satisfied if

$$|G'(x)| \leq \lambda < 1 \text{ for } x \in [a, b]$$

With  $G(x) = \alpha f(x) + x$

$$G'(x) = \alpha f'(x) + 1$$

and we want  $|\alpha f'(x) + 1| \leq \lambda < 1$

So  $f'(x)$  needs to be bounded on  $[a, b]$

# Combine rescaling with equivalent functions

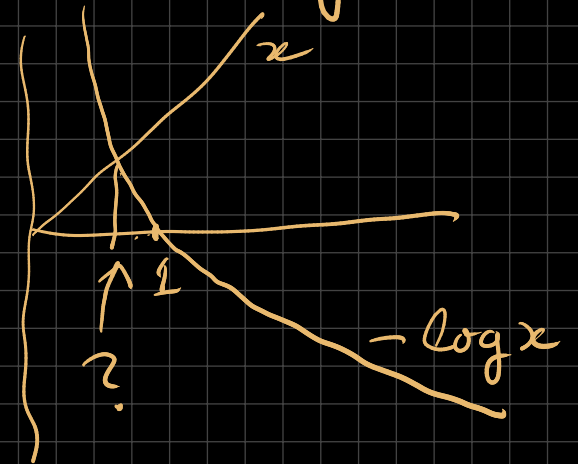
From text p. 33

Problem: Find root of  $f(x) = x + \log x$

Could try:

$$\alpha = -1: \quad G(x) = -(x + \log x) + x \\ = -\log x$$

$$\alpha = -\frac{1}{2}: \quad G(x) = -\frac{1}{2}(x + \log x) + x = \frac{x - \log x}{2}$$



$$\begin{aligned} \textcircled{0} \mathbb{R} \quad x + \log x &= 0 && \text{iff} \quad -x = \log x \\ &&& \text{iff} \quad e^{-x} = x \\ &&& \text{iff} \quad e^{-x} - x = 0 \end{aligned}$$

$$\alpha = \frac{1}{2} : G(x) = \frac{1}{2} (e^{-x} - x) + x = \frac{e^{-x} + x}{2}$$

---

For  $G$  to work we need

$$|G'(x^*)| \leq \lambda < 1 \text{ at solution } x^*$$

to be able to start in an interval

$$[a, b] \ni x^* \in [a, b]$$

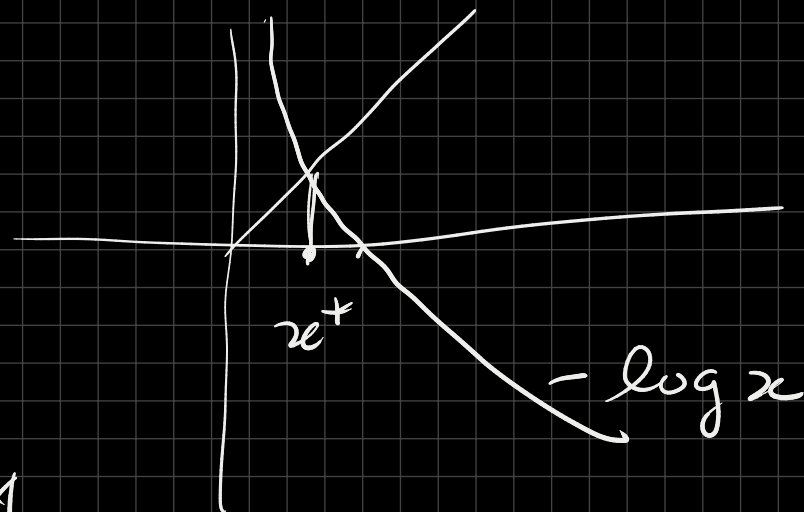
---

Note above:

$$G(x) = -\ln x$$

can't work

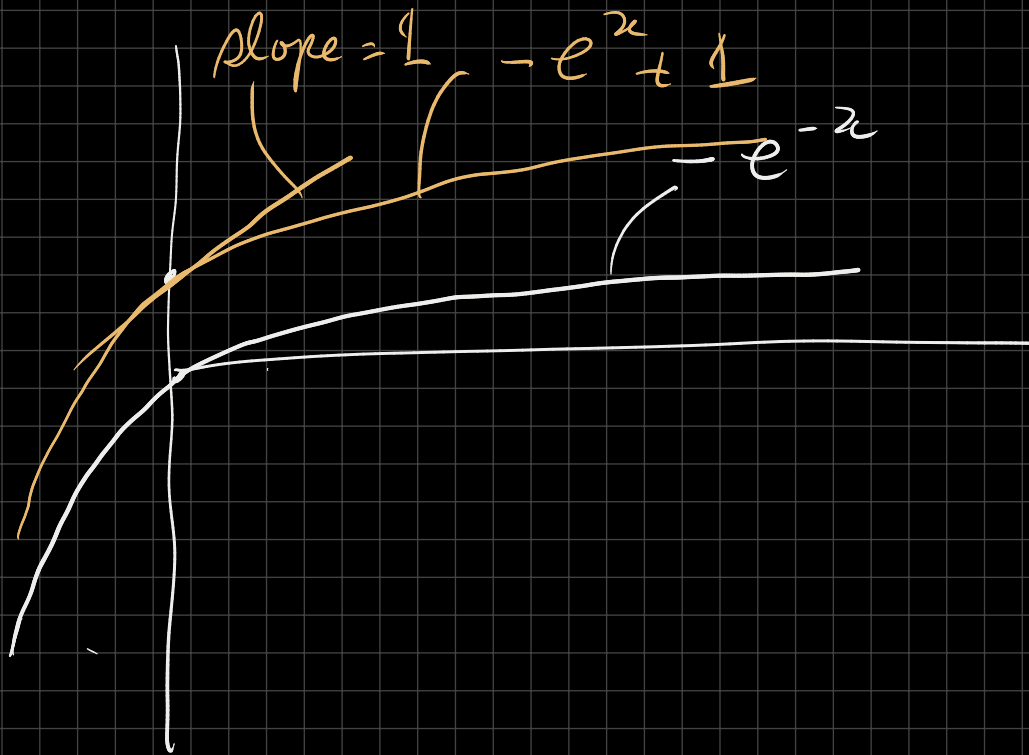
$$\text{since } G'(x) < -1$$



$$G(x) = \frac{e^{-x} + x}{2}$$

$$G'(x) = \frac{-e^{-x} + 1}{2}$$

or  $G'(x) < \frac{1}{2}$   
for  $x > 0$





# Multivariate Problems

Already seen Newton's Method

To max  $g(\underline{x})$

Start at  $\underline{x}_1$ :

Update  $\underline{x}_t$  to  $\underline{x}_{t+1}$

$$\underline{x}_{t+1} = \underline{x}_t - \underbrace{g''(\underline{x}_t)^{-1} g'(\underline{x}_t)}_{\text{increment } h_t}$$

Modifications: Replace  $[g''(x_t)]^{-1}$

For likelihood:

Replace  $g''(\underline{\theta}_{\sim t})$

with  $E_{\underline{\theta}_{\sim t}}(g''(\theta_{\sim t})) = -I(\underline{\theta}_{\sim t})$

so update is  $\underline{h}_{\sim t} = I(\underline{\theta}_{\sim t})^{-1} l'(\underline{\theta}_{\sim t})$

Advantage:  $I(\underline{\theta}_{\sim t})$  is always non-neg. definite  
because

# IRLS : from Generalized Linear Models

Least-squares  $Y = X\beta + \varepsilon$ ,  $E(\varepsilon) = 0$   
 $\text{Var}(\varepsilon) = \sigma^2 I$

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad (\text{Terrible algorithm in finite precision})$$

Better: Use SVD :  $X = U D V^T$

$$U^T U = I, \quad V^T V = I$$

$$D = \begin{bmatrix} d_1 & & 0 \\ & \ddots & \\ 0 & & d_p \end{bmatrix} \quad d_i > 0$$

$$\hat{\beta} = V D^{-1} U^T Y$$

GLM Scaled exponential family

Density or prob. fcn

$$f(y | \theta, \phi) = \exp\left\{ \frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi) \right\}$$

$\theta$  is canonical parameter  
 $\phi$  is dispersion parameter

$$\mu = E_{\theta, \phi}(Y) = b'(\theta) \quad \text{depends only on } \theta$$

$$\text{Var}_{\theta, \phi}(Y) = b''(\theta) a(\phi)$$

Link function  $g(\mu)$  is inverse of  $b'$

i.e solve for  $\theta$  in  $\mu = b'(\theta)$

to get  $\theta = g(\mu)$

Link function

---

## Regression using GLMs

Observe : with <sup>unknown</sup> means Predictors

$y_1$

$\mu_1$

$x_{11}$

...

$x_{p1}$

$y_2$

$\mu_2$

$x_{12}$

...

$x_{p2}$

⋮

⋮

⋮

⋮

$y_n$

$\mu_n$

$x_{1n}$

...

$x_{pn}$

GLM: with canonical link

$$\theta_i = g(\mu_i) = x_{1i} \beta_1 + \dots + x_{pi} \beta_p$$

$$g(\underline{\mu}) = \underline{\theta} = X \underline{\beta}$$

---


Consider  $l(\underline{\theta})$ : treat  $\psi$  as a known parameter

$$n=1 \quad l(\theta) = \frac{y\theta - b(\theta)}{a(\psi)} + k$$

$$l'(\theta) = \frac{y - b'(\theta)}{a(\eta)}$$

Hence  
 $E_{\theta}(Y) = b'(\theta)$

$$l''(\theta) = \frac{-b''(\theta)}{a(\eta)}$$

$$I(\theta) = E_{\theta}(\underbrace{-l''(\theta)}_{\text{not even random!!}}) = \frac{b''(\theta)}{a(\eta)} = \underbrace{-l''(\theta)}$$


$$\begin{aligned} \text{So } l(\underline{\theta}_i) &= \sum_{i=1}^n \frac{y_i \theta_i - b(\theta_i)}{a(\psi)} \\ &= \sum_{i=1}^n y_i \underline{x}_i^T \underline{\beta} - b(\theta_i) \\ &= \underline{y}^T X \underline{\beta} - \sum_{i=1}^n b(\theta_i) \end{aligned}$$



Express  $\underline{l}$  as a function of  $\underline{\beta}$

$$l(\underline{\beta}) = \underline{y}^T X \underline{\beta} + \sum b(\theta_i(\underline{\beta}))$$

Take gradient:

$$l'(\underline{\beta})^T = \underline{y}^T X + \sum \left\{ \overbrace{b'(\theta_i) \theta_i'(\underline{\beta})}^{\text{chain rule}} \right\}$$

$$= \underline{y}^T X + \underline{\mu}^T X$$

$$= (\underline{y} - \underline{\mu})^T X$$

$$e'(\underline{\beta}) = X^T (\underline{y} - \underline{\mu})$$

$$e''(\underline{\beta}) = -X^T \frac{d\underline{\mu}}{d\underline{\beta}}$$

$$\begin{cases} b''(\theta_i) & i=k \\ 0 & i \neq k \end{cases}$$

now

$$\frac{d\underline{\mu}}{d\underline{\beta}} = \begin{bmatrix} \frac{d\mu_i}{d\beta_j} \end{bmatrix} = \begin{bmatrix} \sum_{k=1} \frac{d\mu_i}{d\theta_k} \frac{d\theta_k}{d\beta_j} \end{bmatrix}$$

$$= \begin{bmatrix} b''(\theta_1) & & 0 \\ & \ddots & \\ 0 & & b''(\theta_n) \end{bmatrix} \times X$$

since  $\mu = b'(\theta)$   
 $\frac{d\mu}{d\theta} = b''(\theta)$

$x_{kj}$  from  $k^{\text{th}}$  row and  $j^{\text{th}}$  col of  $X$

$$\text{So } \ell''(\underline{\beta}) = -X^T \underbrace{W}_\text{diagonal matrix} X$$

$$\text{and } \underline{\beta}_{t+1} = \underline{\beta}_t + (X^T W_t X)^{-1} X^T (y - \underline{\mu}_t)$$

$$\underline{\theta}_{t+1} = X \underline{\beta}_{t+1}$$

$$\underline{\mu}_{t+1} = b'(\underline{\theta}_{t+1})$$

$$W_{t+1} = \text{diag}(b''(\underline{\theta}_{t+1}))$$

$\underline{\tilde{y}} - \underline{\tilde{\mu}}_t$  from the last step is the  
vector of response residuals.

That's the general form.

Let's apply it to a logistic regression

$$P(y) = \pi^y (1-\pi)^{1-y} \quad y \in \{0, 1\}$$

$$= \exp\{y \ln \pi + (1-y) \ln(1-\pi)\}$$

$$= \exp\left\{y \ln\left(\frac{\pi}{1-\pi}\right) + \ln(1-\pi)\right\}$$

$$\text{So } b(\pi) = -\ln(1-\pi), \quad a(\phi) \equiv \underline{1}$$

$$\theta = \ln\left(\frac{\pi}{1-\pi}\right) \quad \pi = \frac{e^\theta}{1+e^\theta}$$

$$\begin{aligned} b(\theta) &= -\ln\left(1 - \frac{e^\theta}{1+e^\theta}\right) = -\ln\left(\frac{1}{1+e^\theta}\right) \\ &= \ln(1+e^\theta) \end{aligned}$$

$$b'(\theta) = \frac{e^\theta}{1+e^\theta} = \pi = \frac{1}{1+e^{-\theta}}$$

$$b''(\theta) = \frac{e^\theta(1+e^\theta) - e^{2\theta}}{(1+e^\theta)^2} = \text{expit}(\theta)$$

$$= \frac{e^\theta}{(1+e^\theta)^2} = \pi(1-\pi)$$

## Recall

$$\tilde{\beta}_{t+1} = \tilde{\beta}_t + (X^T W_t X)^{-1} X^T (y - \tilde{\mu}_t)$$

$$\tilde{\theta}_{t+1} = X \tilde{\beta}_{t+1}$$

$$\tilde{\mu}_{t+1} = b'(\tilde{\theta}_{t+1})$$

$$W_{t+1} = \text{diag}(b''(\tilde{\theta}_{t+1}))$$

$$\tilde{\beta}_{t+1} = \tilde{\beta}_t + (X^T W_t X)^{-1} X^T (y - \tilde{\mu}_t)$$

$$\tilde{\theta}_{t+1} = X \tilde{\beta}_{t+1}$$

$$\tilde{\mu}_{t+1} = \tilde{\pi}_{t+1} = \text{expit}(\tilde{\theta}_{t+1})$$

$$W_{t+1} = \text{diag}(\tilde{\pi}_{t+1} (1 - \tilde{\pi}_{t+1}))$$

---

Cross-check with text p.

# Newton-like methods

Newton works in one step  
if  $g$  is quadratic because

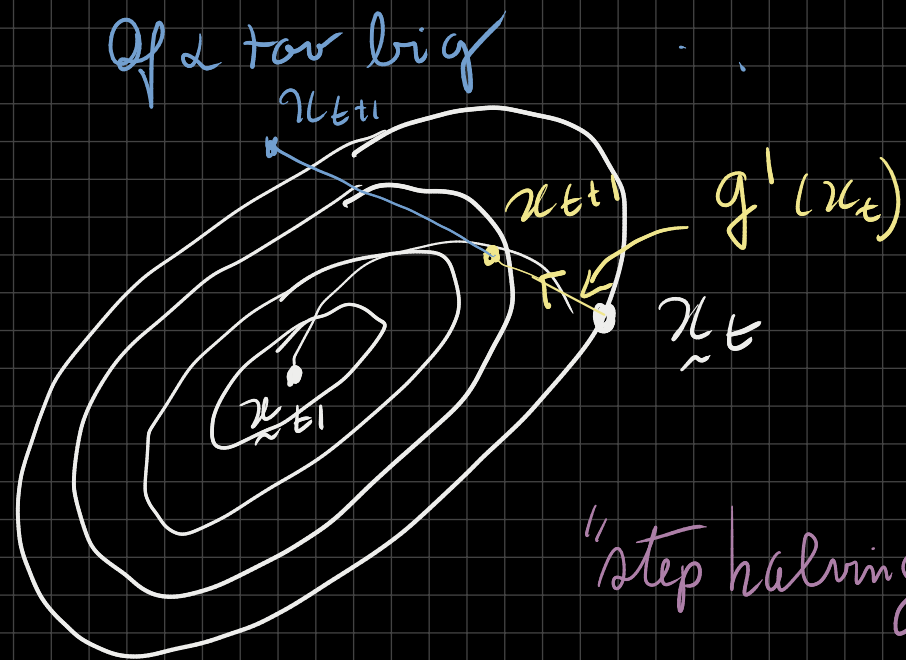
$$\tilde{x}_{t+1} = \tilde{x}_t + \underbrace{\left[-g''(\tilde{x}_t)\right]^{-1} g'(\tilde{x}_t)}_{\text{knows exactly how far to go.}}$$

But if not quadratic ???



Instead of  $x_{\tilde{t}+1} = x_{\tilde{t}} + [-g''(x_{\tilde{t}})]^{-1} g'(x_{\tilde{t}})$

How about a small step up:  $x_{\tilde{t}+1} = x_{\tilde{t}} + \alpha \mathbb{I} g'(x_{\tilde{t}})$



Choosing  $\alpha$ :

Try some  $\alpha = \alpha_0$

If  $g(x_{\tilde{t}+1}) < g(x_{\tilde{t}})$

then set  $\alpha = \alpha/2$  & repeat.

Instead of  $I$  use something else

e.g.  $A = [-g''(\underline{x}_0)]^{-1}$

Only works if  $-g''(\underline{x}_0)$  is positive-definite

Lots of variants go by acronyms using names of people who published them.

e.g. BFGS

Broyden - Fletcher - Goldfarb - Shanno

Davidon - Fletcher - Powell (DFP)

Gauss-Newton:

Non-linear least squares

Linear LS

$$\underline{y} = X \underline{\beta} + \underline{\epsilon}$$

$$E(\underline{\epsilon}) = \underline{0} \quad \text{Var}(\underline{\epsilon}) = \sigma^2 W$$

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

terrible algorithm,

Non-linear

$$y_i = f(\underline{x}_i, \underline{\theta}) + \varepsilon_i$$

Use a linear approximation for  $y_i$

$$y_i \approx f(\underline{x}_i, \underline{\theta}_t) + f'(\underline{x}_i, \underline{\theta}_t) (\underline{\theta} - \underline{\theta}_t) + \varepsilon_i$$

i.e.  $y_i - f(\underline{x}_i, \underline{\theta}_t) \approx f'(\underline{x}_i, \underline{\theta}_t) (\underline{\theta} - \underline{\theta}_t) + \varepsilon_i$

Do LS reg of " $y_t$ " on  $X_t \eta_t$

and set  $\underline{\theta}_{t+1} = \underline{\theta}_t + \eta_t$



