Convergence Order

How fast does $|\varepsilon_t| \to 0$?

Convergence order β if

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

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

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

E.g. if $\epsilon_t = .001$ $\beta = 1$ $c = 1/2$

then $\epsilon_{t+1} \approx .0005$

$\epsilon_{t+2} \approx .00025$ ...

≈ 2.5 steps to get to 1.0001

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

$\epsilon_{t+1} = .00000025$

etc.

Fisher Scoring

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

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

is at least pos. semi-definite.

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

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

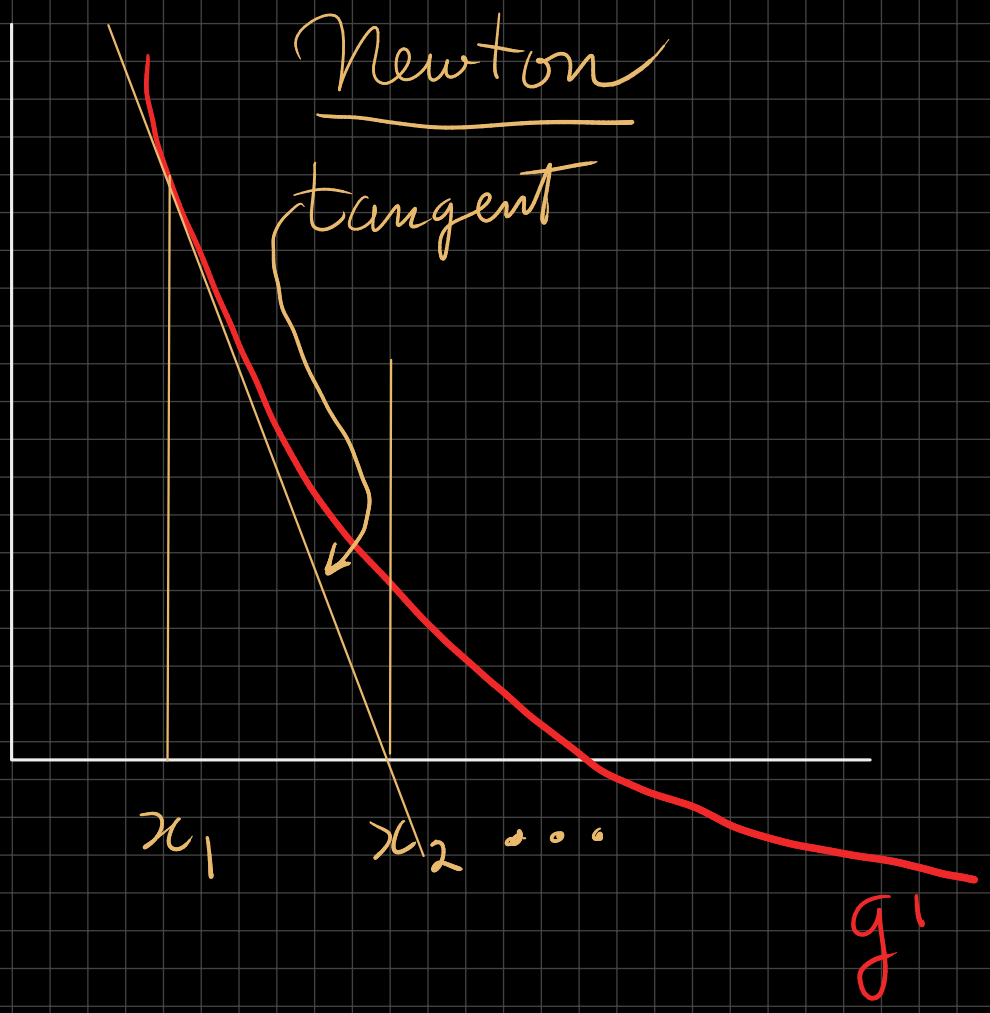
Secant Method: For $\theta \in \mathbb{R}$

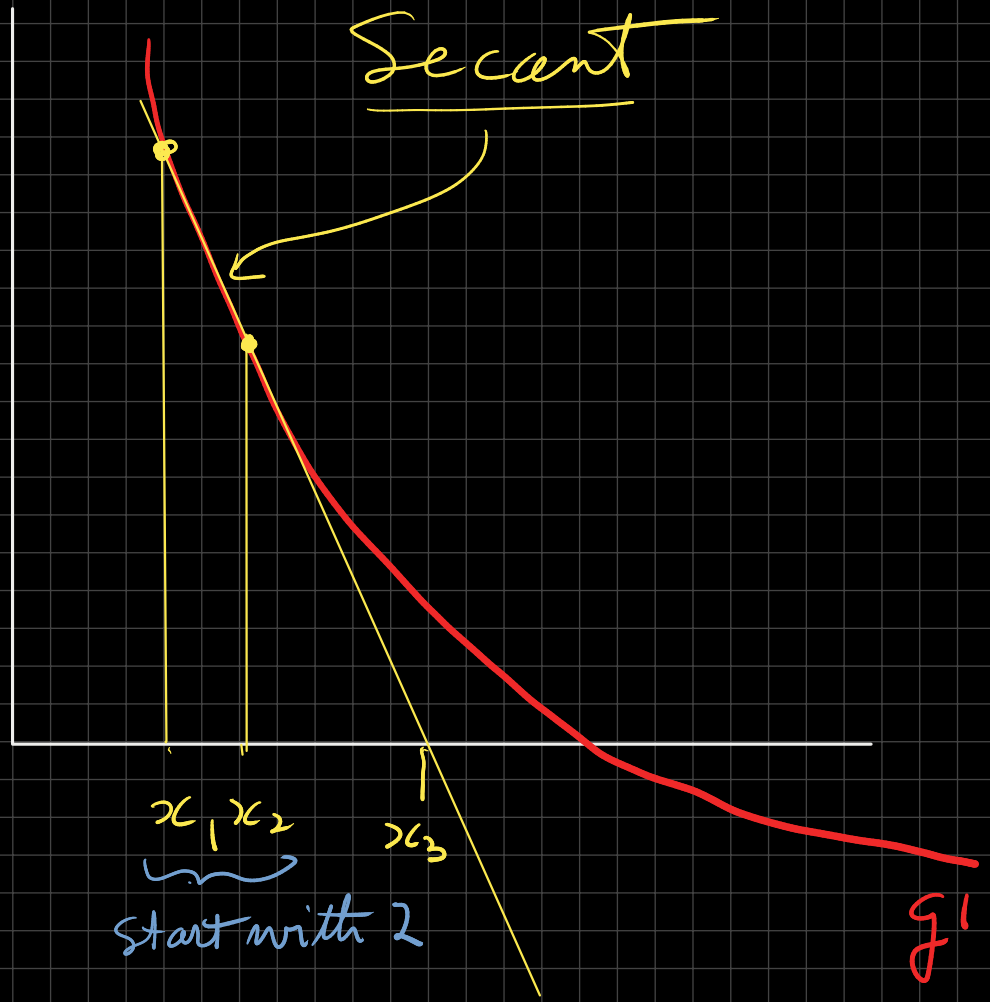
Use $$\frac{|g'(\theta_t) - g'(\theta_{t-1})|}{\theta_t - \theta_{t-1}}$$

To guess $l''(\theta_t)$

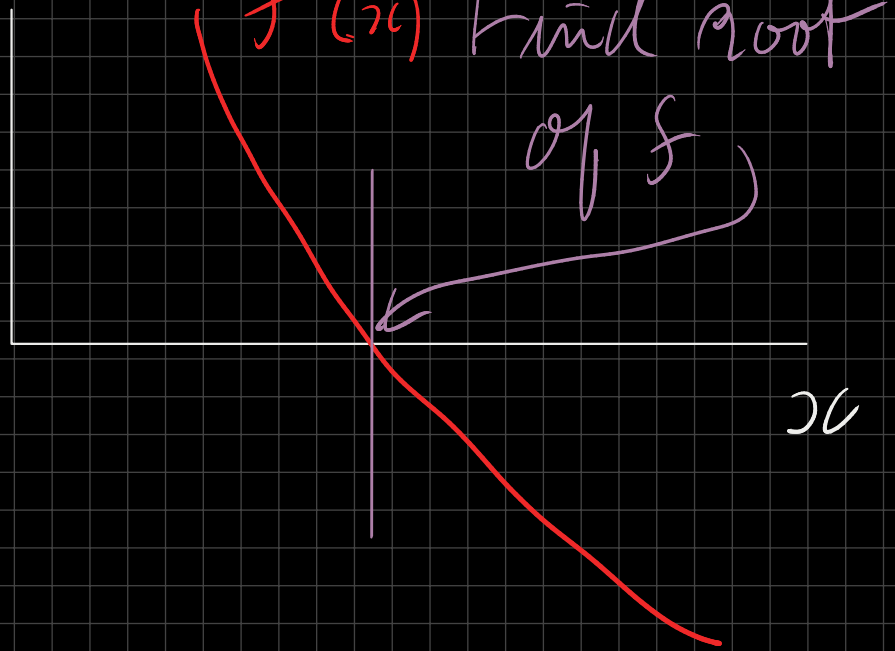
- Need 2 starting values. θ_1, θ_2
- Will adapt for $\mathbb{R}^p$

Why "secant"

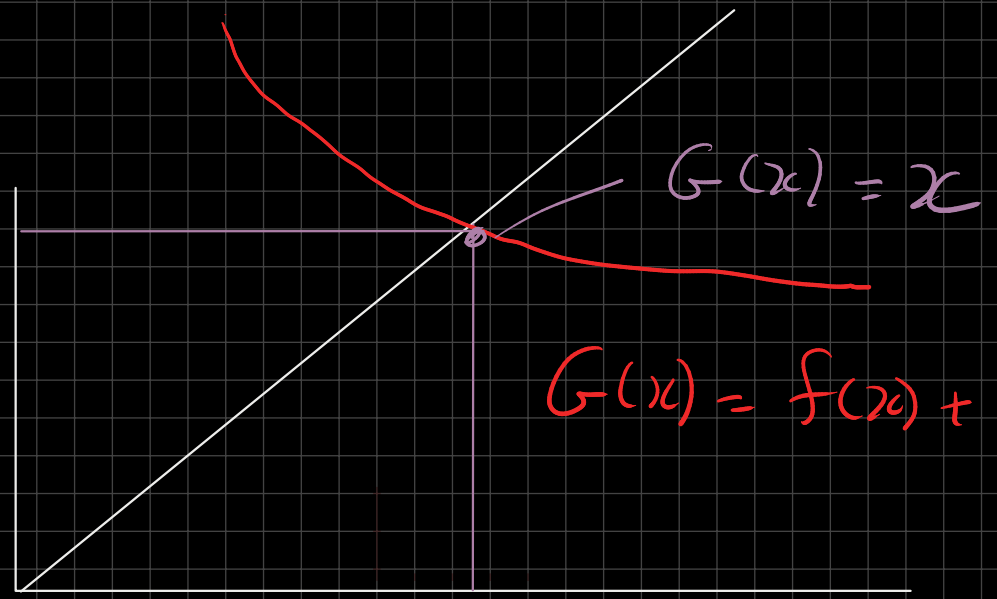




Fixed point iteration



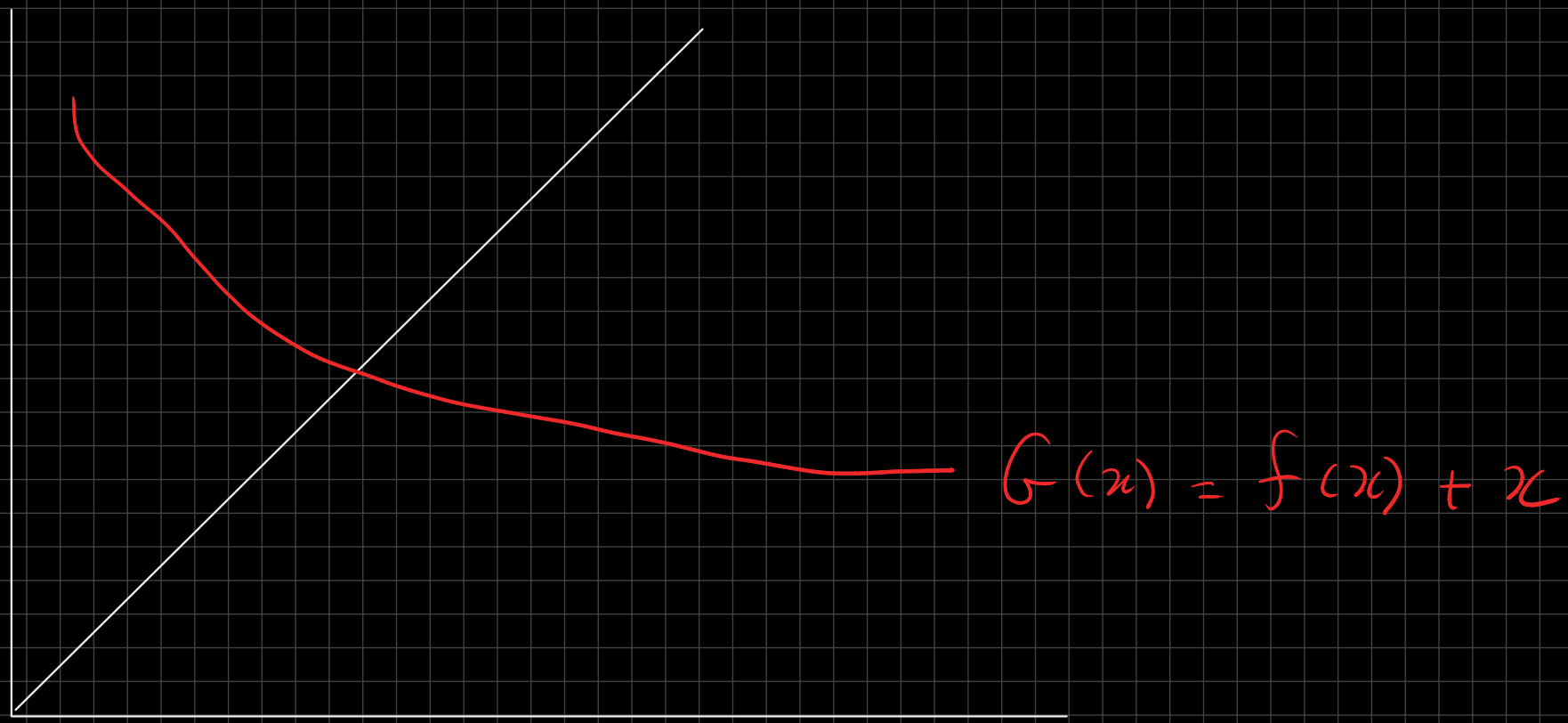
same as



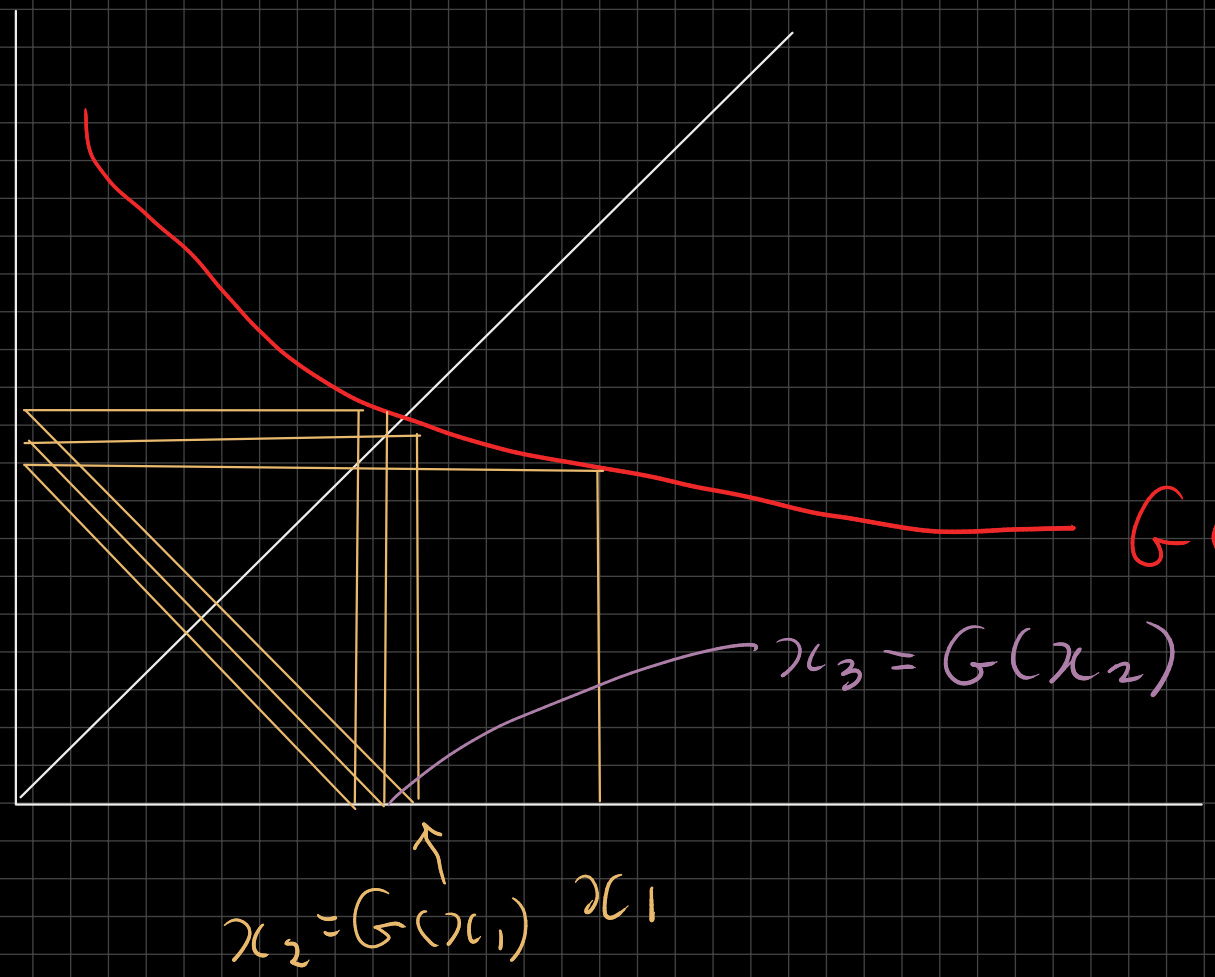
Idea: $x_{t+1} = G(x_t)$

until $x_{t+1} \simeq x_t$

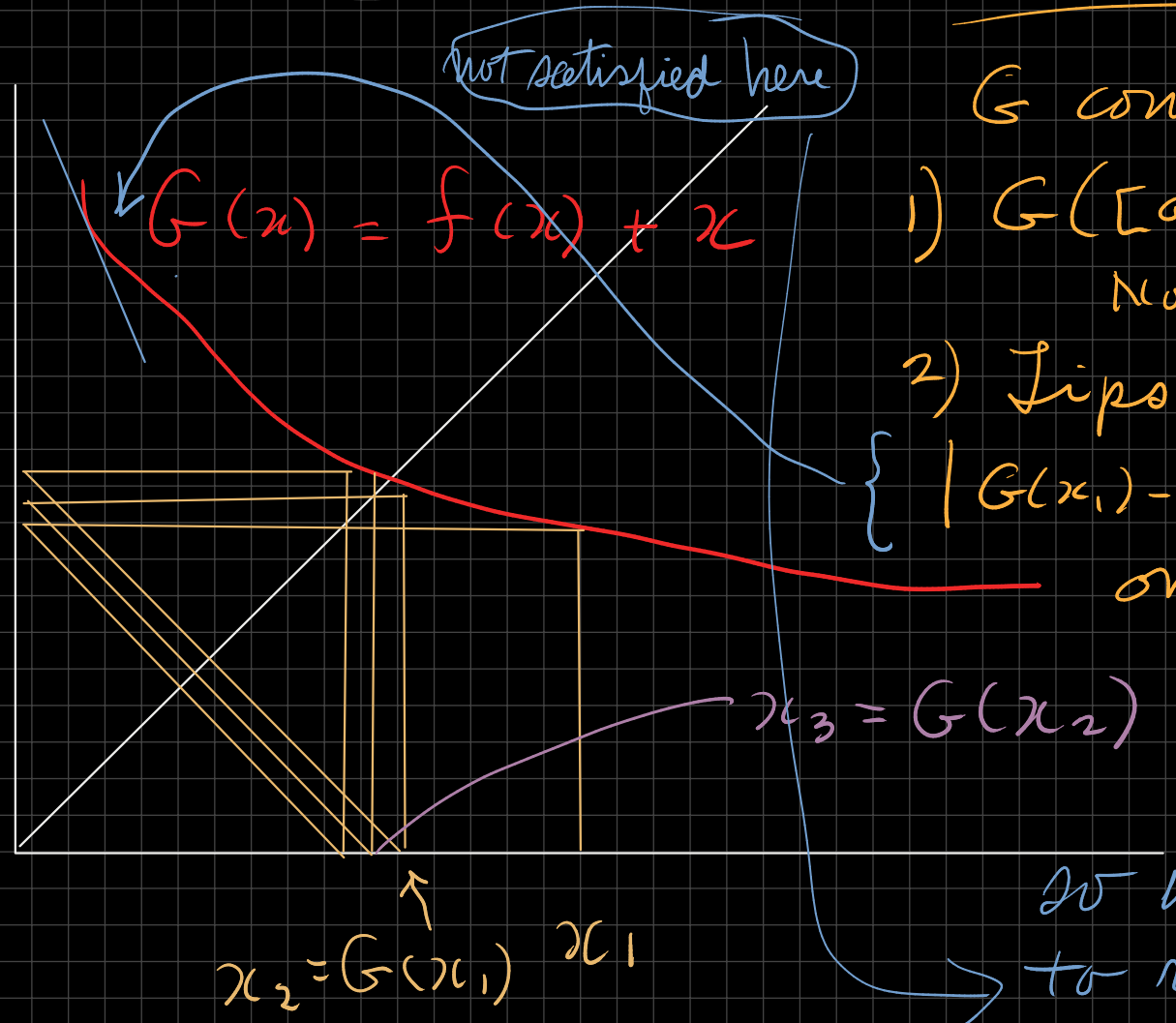
How does this work?



How does this work?



How does this work?



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

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

$|G(x_1) - G(x_2)| \leq \lambda |x_1 - x_2|$ on $[a,b]$ for some $\lambda \in [0,1)$

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

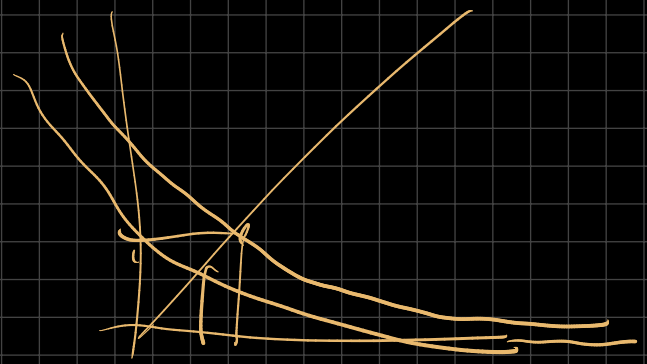
So $f'(x)$ needs to be bounded on $[a, b]$

Other strategy: Find a equivalent problem by e.g. reparametrizing:

Maximize $g(x) = \dfrac{\ln x}{1+x}$

Find a root of $f(x) = g'(x) = \dfrac{\frac{1}{x} - \ln x}{(1+x)^2}$

$\forall x \neq -1$, $f(x) = 0$ iff $1 - x \ln x = 0$



$$f(x) = x + \ln x$$

$$\ln x = -x$$

iff $x = e^{-x}$

$$e^{-x} - x = 0$$

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

$\alpha = 1$

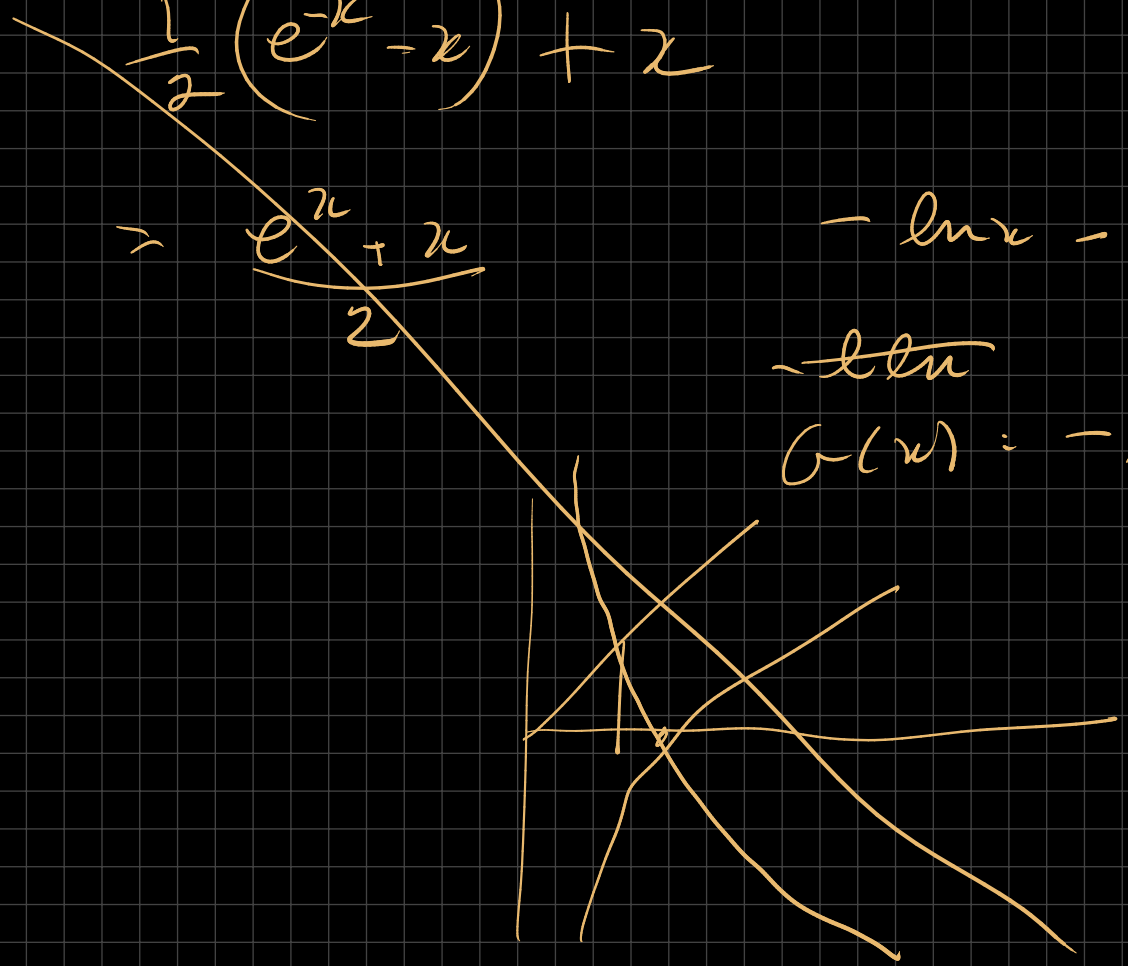
$$\boxed{f(x) + x = e^{-x}}$$

$$\frac{1}{2}\left(e^{-x} - x\right) + x$$

$$= \frac{e^{x} + x}{2}$$

$$-\ln x - x = 0$$

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

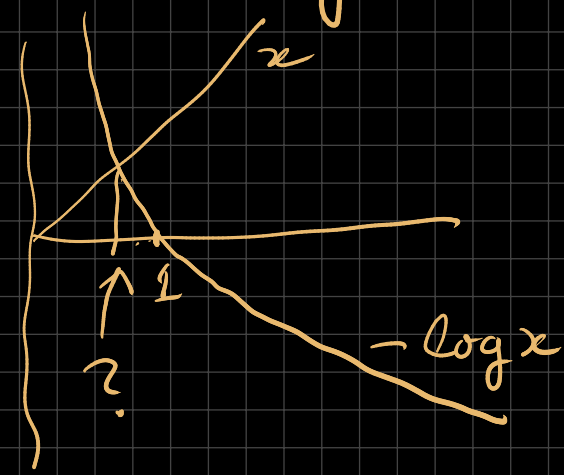


Combine rescaling with equivalent functions

From text p. 33

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

Could try:



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

$= -\log x$

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

OR $x + \log x = 0$ iff $-x = \log x$

iff $e^{-x} = x$

iff $e^{-x} - x = 0$

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

For G to work we need

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

to be able to start in an interval

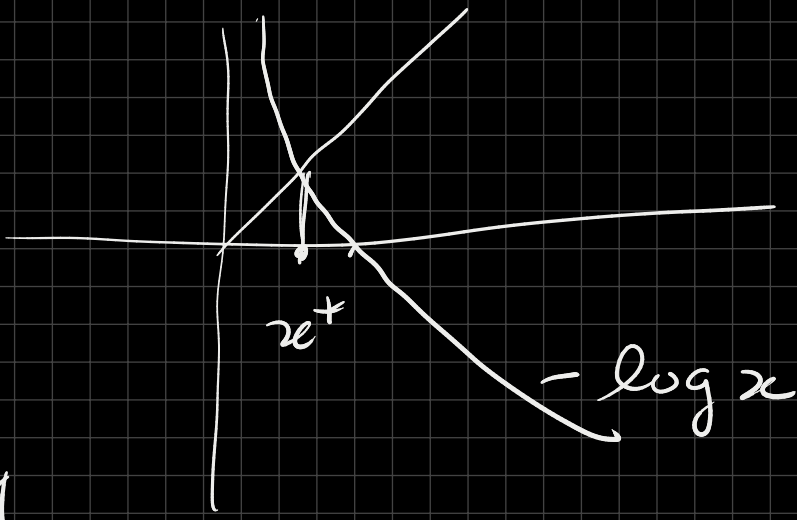
$$[a, b] \ni x^* \in [a, b]$$

Note above:

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

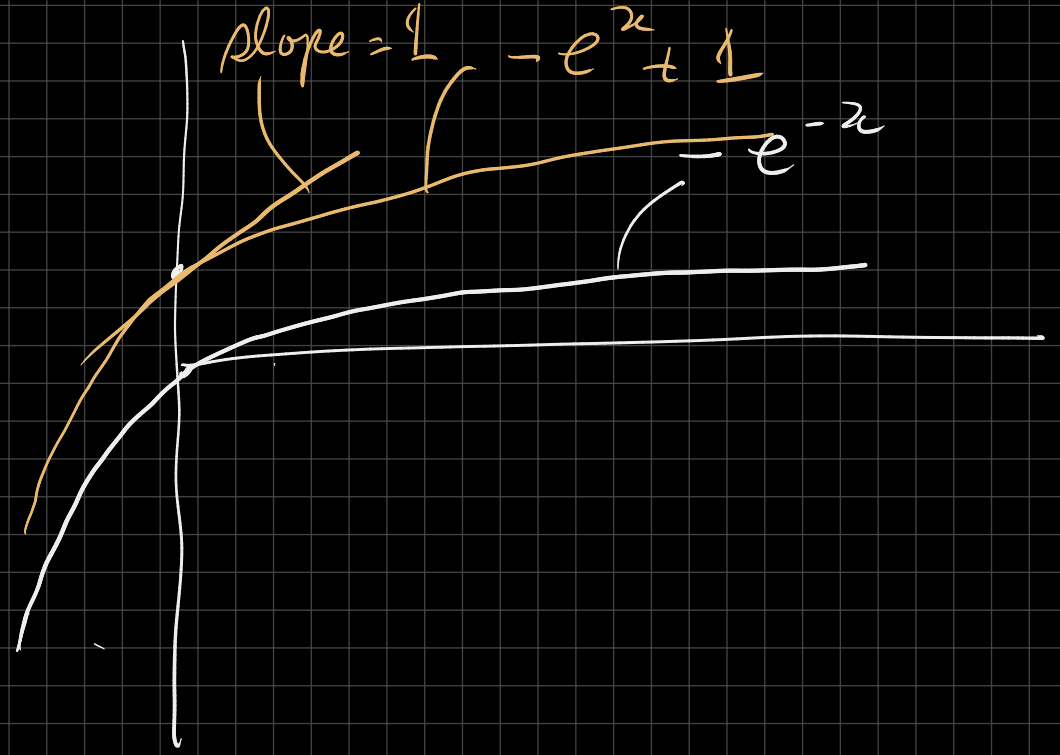
can't work

since $G'(x) < -1$



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

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



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

for $x > 0$

Multivariate Problems

Already seen Newton's Method

To max $g(\underset{\sim}{x})$

Start at $\underset{\sim}{x}_1$:

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

$$\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_x - \underbrace{g''(\underset{\sim}{x}_t)^{-1} g'(\underset{\sim}{x}_t)}_{\text{increment } \underset{\sim}{h}_t}$$

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

For likelihood:

Replace $g''(\underset{\sim}{\theta}_t)$

with $E_{\underset{\sim}{\theta}_t}(g''(\theta_t)) = -I(\underset{\sim}{\theta}_t)$

So update is $\underset{\sim}{h}_t = I(\underset{\sim}{\theta}_t)^{-1} l'(\underset{\sim}{\theta}_t)$

Advantage: $I(\underset{\sim}{\theta}_t)$ is always non-neg. definite because

IRLS : from Generalized Linear Models

Least-squares $Y = X\beta + \varepsilon$, $E(\varepsilon) = 0$

$Var(\varepsilon) = \sigma^2 I$

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

Better: Use SVD : $X = UDV^T$

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

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

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

GLM

Scaled exponential family

density or prob. fcn

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

θ is canonical parameter

ϕ is dispersion parameter

$$\text{"}\mu\text{"} = \mathop{E}_{\theta,\phi}(Y) = b'(\theta)$$ depends only on θ

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

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

i.e solve for θ in $\mu = b'(\theta)$

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

Regression using GLMs

Observe : with unknown means : Predictors

$$\begin{array}{ccc} y_1 & \mu_1 & x_{11} \cdots x_{p1} \\ y_2 & \mu_2 & x_{12} \cdots x_{p2} \\ \vdots & \vdots & \vdots \\ y_n & \mu_n & x_{1n} \cdots x_{pn} \end{array}$$

GLM : with canonical link

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

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

Consider $l(\underset{\sim}{\theta})$: treat ψ as a known parameter

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

$$\ell'(\theta) = \frac{y - b'(\theta)}{a(\psi)}$$

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

$$\ell''(\theta) = \frac{-b''(\theta)}{a(\psi)}$$

$$I(\theta) = E_\theta(-\ell''(\theta)) = \frac{b''(\theta)}{a(\psi)} = \underline{-\ell''(\theta)}$$

not even random!!

So $$l(\underset{\sim}{\theta}_i) = \sum_{i=1}^{n} \frac{y_i \theta_i - b(\theta_i)}{a(\psi)}$$

$$= \sum_{i=1}^{n} y_i \underset{\sim}{x}_i^T \underset{\sim}{\beta} - b(\theta_i)$$

$$= \underset{\sim}{y}^T X \underset{\sim}{\beta} - \sum_{i=1}^{n} b(\theta_i)$$

Express $\underset{\sim}{l}$ as a function of $\underset{\sim}{\beta}$

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

Take gradient:

$$l'(\underset{\sim}{\beta})^T = \underset{\sim}{y}^T X + \sum \overbrace{\left\{ b'(\theta_i)\, \theta_i'(\underset{\sim}{\beta}) \right\}}^{\text{Chain rule}}$$

$$= \underset{\sim}{y}^T X + \underset{\sim}{\mu}^T X$$

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

$$\ell'(\underset{\sim}{\beta}) = X^T(\underset{\sim}{y} - \underset{\sim}{\mu})$$

$$\ell''(\underset{\sim}{\beta}) = -X^T \frac{d\underset{\sim}{\mu}}{d\underset{\sim}{\beta}}$$

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

since $\mu = b'(\theta)$

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

now $\frac{d\underset{\sim}{\mu}}{d\underset{\sim}{\beta}} = \left[\frac{d\mu_i}{d\beta_j}\right] = \left[\sum_{k=1} \frac{d\mu_i}{d\theta_k} \frac{d\theta_k}{d\beta_j}\right]$

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

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

So $l''(\underset{\sim}{\beta}) = -X^T \underset{\sim}{W} X$

$\underset{\sim}{W}$: diagonal matrix

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

$$\underset{\sim}{\theta}_{t+1} = \underset{\sim}{X}\,\underset{\sim}{\beta}_{t+1}$$

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

$$W_{t+1} = \mathrm{diag}\left(b''(\underset{\sim}{\theta}_{t+1})\right)$$

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

That's the general form.

Let's apply it to a logistic regression

$$p(y) = \pi^y (1-\pi)^{1-y} \qquad y \in \{0, 1\}$$

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

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

So $b(\pi) = -\ln(1-\pi)$; $a(\phi) = \underline{1}$

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

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

$$= \ln(1+e^\theta)$$

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

$$= \text{expit}(\theta)$$

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

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

Recall

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

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

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

$$W_{t+1} = \text{diag}\left(b''(\underset{\sim}{\theta}_{t+1})\right)$$

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

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

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

$$W_{t+1} = \text{diag}\left(\underset{\sim}{\pi}_{t+1}(1 - \underset{\sim}{\pi}_{t+1})\right)$$

Cross-check with text p.

Newton-like methods

Newton works in one step if g is quadratic because

$$\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t + \underbrace{\left[-g''(\underset{\sim}{x}_t)\right]^{-1} g'(\underset{\sim}{x}_t)}$$

knows exactly how far to go.

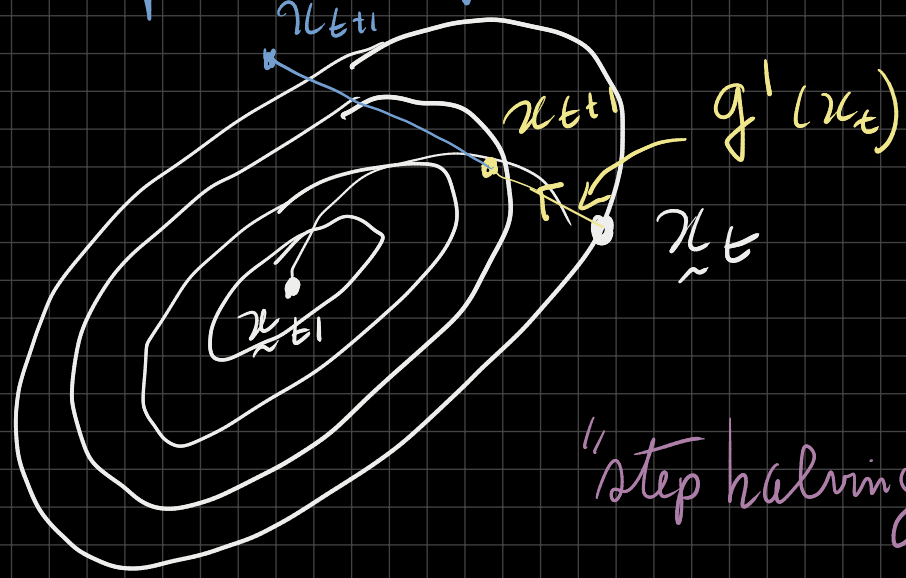
But if <u>not</u> quadratic ???

Instead of $\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t + [-g''(\underset{\sim}{x}_t)]^{-1} g'(\underset{\sim}{x}_t)$

How about a small step up: $\underset{\sim}{x}_{t+1} = \underset{\sim}{x}_t + \alpha \, \underline{I} \, g'(\underset{\sim}{x}_t)$

step size

If α too big



"step halving"

Choosing α:

Try some $\alpha = \alpha_0$

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

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

Instead of I use something else

e.g. $A = \left[-g''(x_0)\right]^{-1}$

Only works if $-g''(x_0)$ is positive-definite

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

eg BFGS

Broyden - Fletcher - Goldfarb - Shanno

Davidon - Fletcher - Powell (DFP)

Gauss-Newton:

Non-linear least squares

Linear LS $\quad \underset{\sim}{Y} = X\underset{\sim}{\beta} + \underset{\sim}{\varepsilon}$

$E(\underset{\sim}{\varepsilon}) = \underset{\sim}{0} \quad Var(\underset{\sim}{\varepsilon}) = \sigma^2 W$

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

terrible algorithm.

Non-linear

$$Y_i = f(\underset{\sim}{x}_i, \underset{\sim}{\theta}) + \varepsilon_i$$

Use a linear approximation for Y_i

$$Y_i \approx f(\underset{\sim}{x}_i, \underset{\sim}{\theta}_t) + f'(\underset{\sim}{x}_i, \underset{\sim}{\theta}_t)(\underset{\sim}{\theta} - \underset{\sim}{\theta}_t) + \varepsilon_i$$

i.e. $Y_i - f(\underset{\sim}{x}_i, \underset{\sim}{\theta}_t) \approx f'(x_i, \theta_t)(\underline{\theta} - \underline{\theta}_t) + \varepsilon_i$

Do LS reg of "Y_t" on X_t

and set $\theta_{t+1} = \theta_t + \eta_t$