Chapter 2

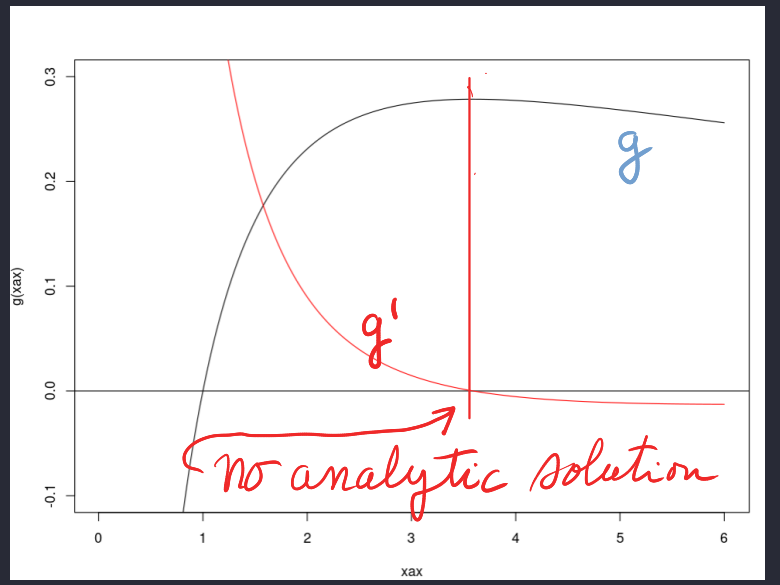
- Optimizing $f : \mathbb{R}^p \to \mathbb{R}$
  where f smooth - differentiable
- Frequent application:
  - Optimize log-likelihood to get MLE.
  - e.g. Find solution, $\hat{\underset{\sim}{\theta}}$, to
    the score equation $l'(\hat{\underset{\sim}{\theta}}) = \underset{\sim}{0}$ } root problem

Univariate ⊖

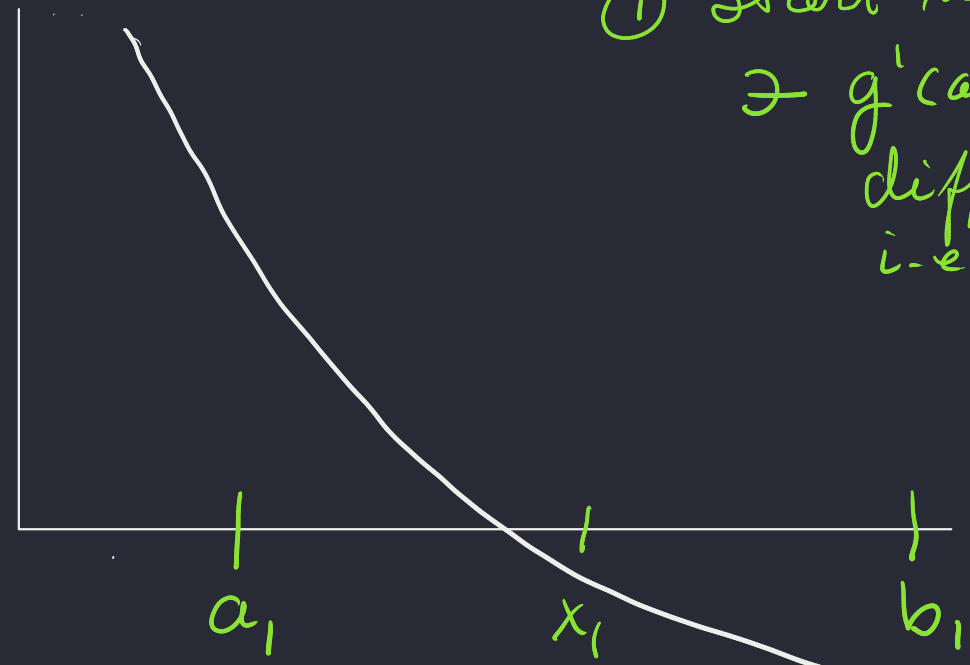
Example:

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

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



Bisection method : Use g'



① Start with a_1, b_1
such $g'(a_1) \cdot g'(b_1)$ have different signs
i.e $g'(a_1) \cdot g'(b_1) < 0$

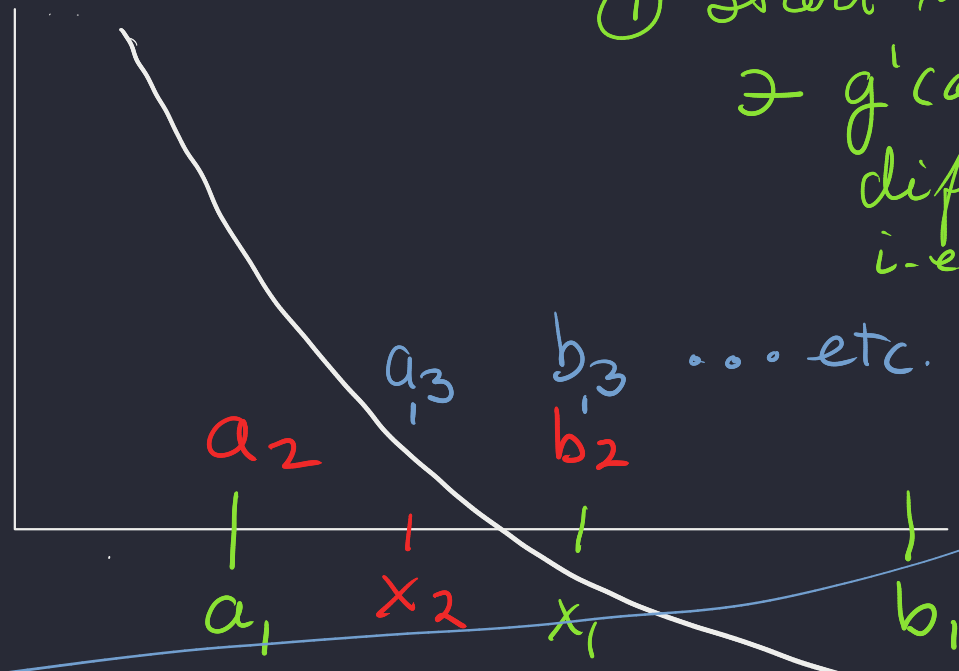
② Bisect
$x_1 = (a_1 + b_1)/2$

③ Check
$g(x_1) \cdot g(b_1)$

If < 0 let $a_2 = x_1$, $b_2 = b_1$
if > 0 let $a_2 = a_1$, $b_2 = x_1$

Repeat.

Bisection method : Use g'



① Start with a_1, b_1
s.t. $g'(a_1)$ & $g'(b_1)$ have different signs
i.e $g'(a_1) \cdot g'(b_1) < 0$

... etc.

② Bisect
$x_1 = (a_1 + b_1)/2$

③ Check
$g(x_1) \cdot g(b_1)$

If < 0 let $a_2 = x_1$, $b_2 = b_1$
if > 0 let $a_2 = a_1$, $b_2 = x_1$
Repeat.

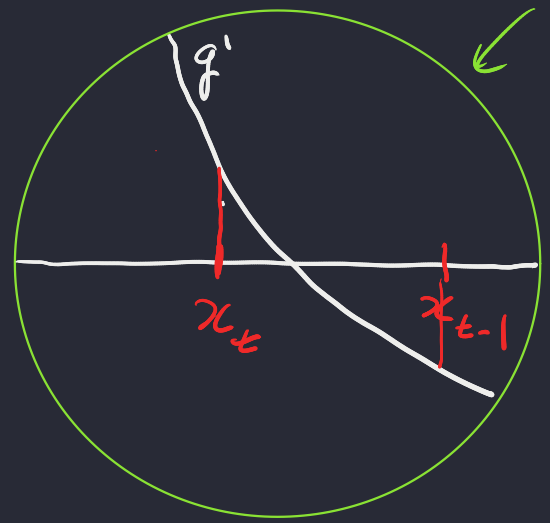
But code:

$$x_1 = a_1 + \frac{b_1 - a_1}{2}$$

Convergence Criterion:

$|x_t - x_{t-1}|$ small

$|g'(x_t)|$ small



↙ microscope

"Absolute" converge criterion $|x_t - x_{t-1}| < \varepsilon$

- Choose small ε
- Converge if $|x_t - x_{t-1}| < \varepsilon$
- Bisection must converge since $|x_t - x_{t-1}| < 2^{-t}|b_1 - a_1|$

"Relative Convergence"

a) $$\frac{|x_t - x_{t-1}|}{|x_t|} < \varepsilon$$

- Independent of units
- Not good to solution at 0!

b) $$\frac{|x_t - x_{t-1}|}{|x_t| + \varepsilon} < \varepsilon$$

Caveats:- Of many roots, will find one

- not necessarily max
- g' must be continuous

Newton's method

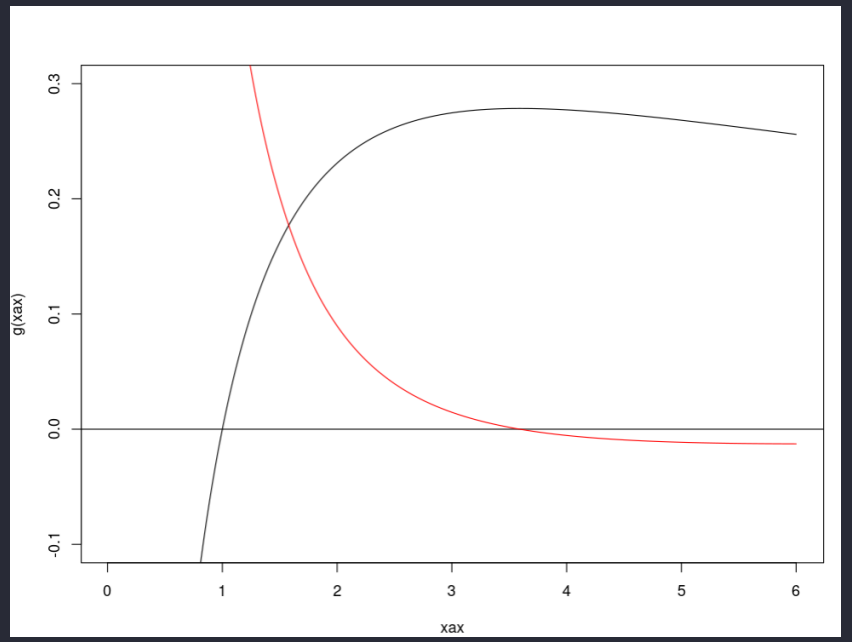
Find x_* such that

$g(x_*)$ is max of $g(x)$

i.e. $x_* = \text{argmax}(g)$

OR Find x_* such that

$g'(x_*) = 0$



i.e 1) max of g by using quad. appx of g

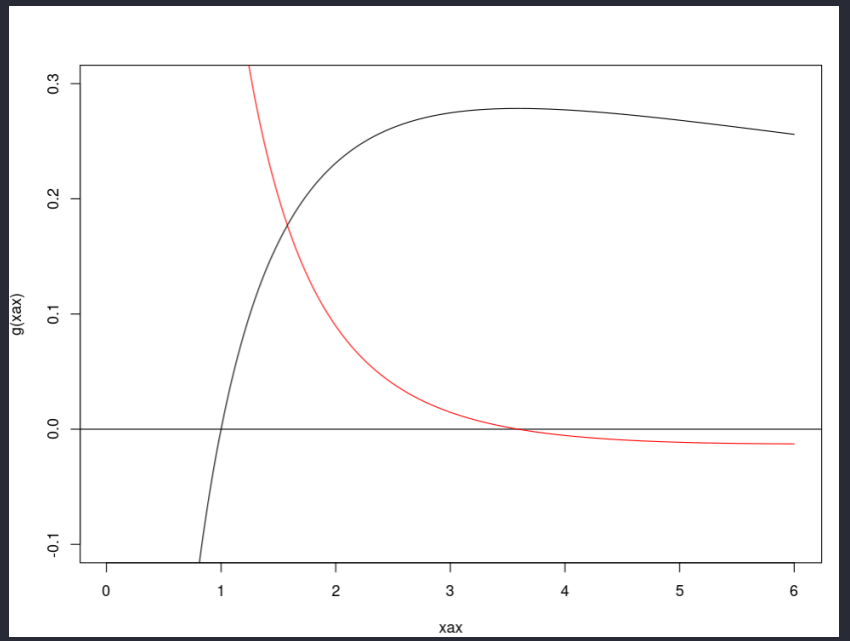
or 2) Root of g' by using a linear appx of g'

Start somewhere

e.g $x_1 = 2$

1) Do a quadratic approximation of g at x_1

2) Solve for max using quad. approx.



OR: 1) Do a linear approximation of g' at at x_1

2) Solve for $\hat{g}'(x) = 0$

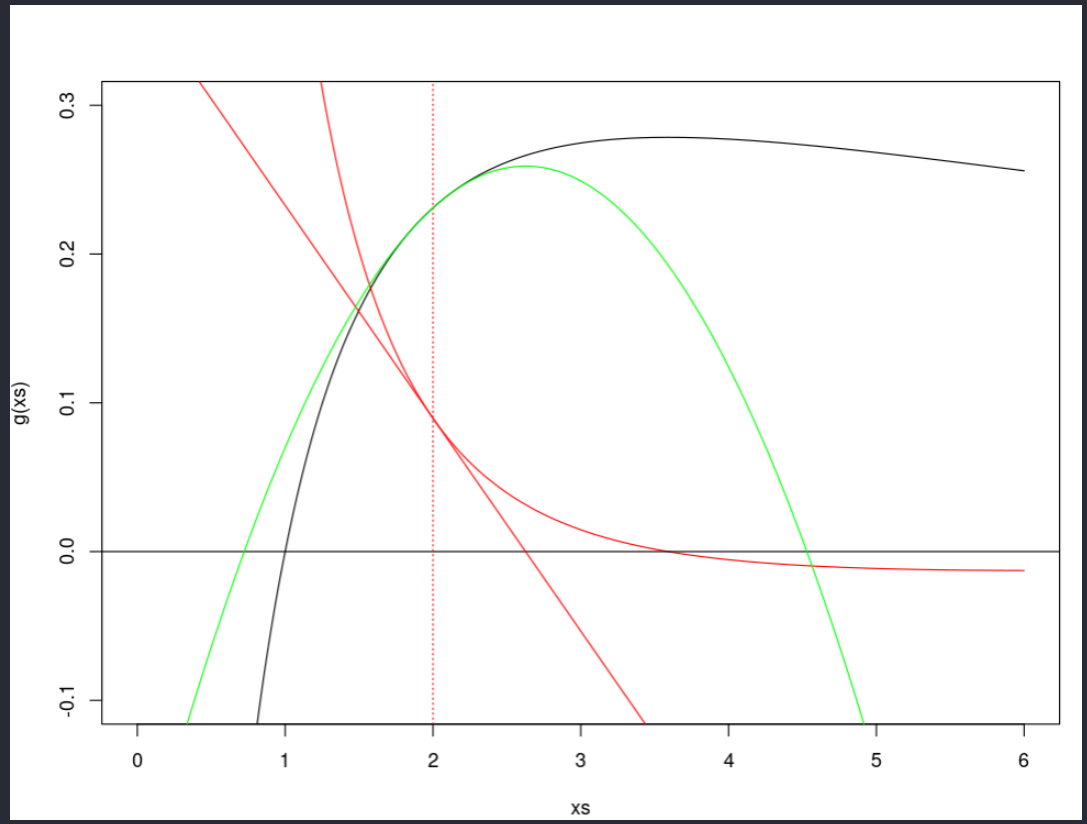
Quadratic approx

$$\hat{g}(x) = g(x_1) + g'(x_1) \times (x - x_1) + g''(x_1) \times (x - x_1)^2 / 2$$

Quad: $a + bx + cx^2$

if $c < 0$ then max at

$$x = -b/2c$$

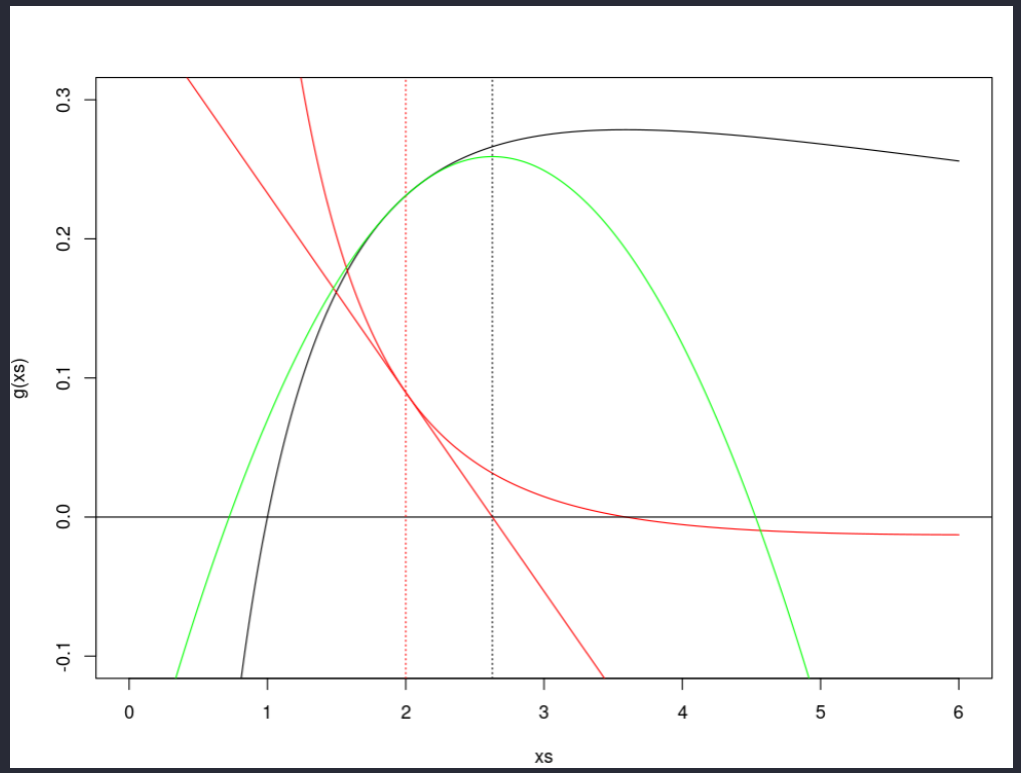


So max of $\hat{g}$ is at $(x - x_1) = -g'(x_1)/g''(x_1)$

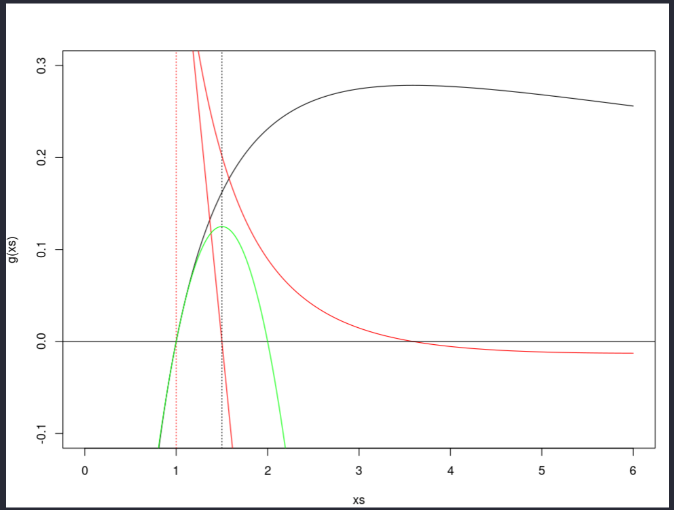
i.e. $x_2 = x_1 - g'(x_1)/g''(x_1)$

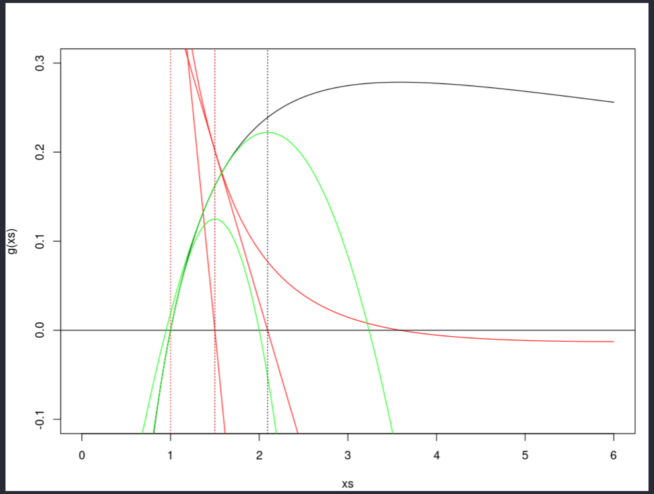
EXER: Show solving for $g'(x) = 0$ yields the same solution.

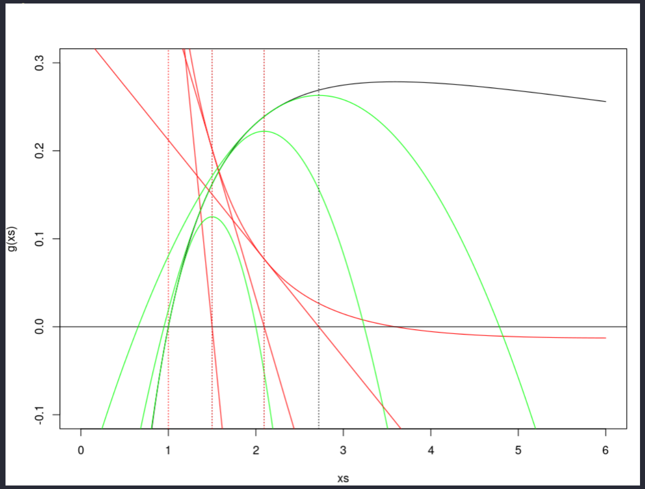
Repeat

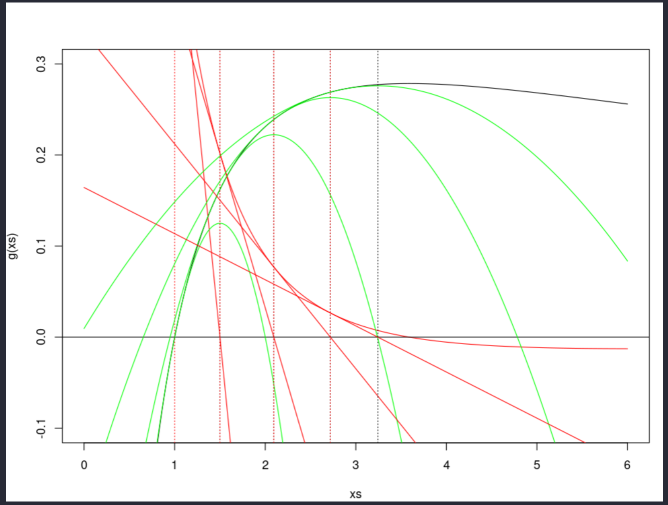


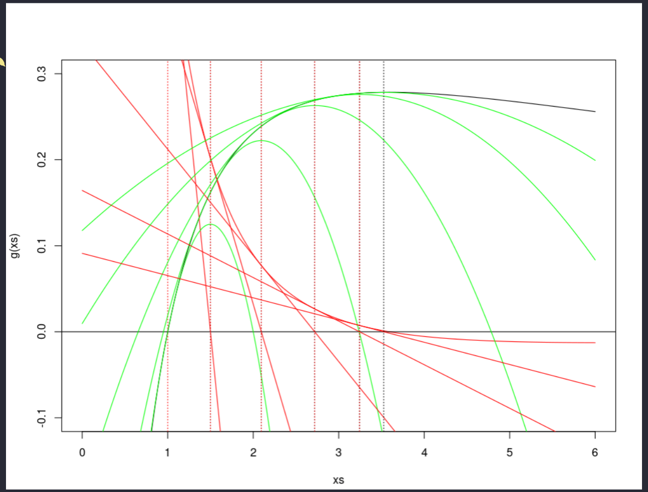
Starting at x = 1

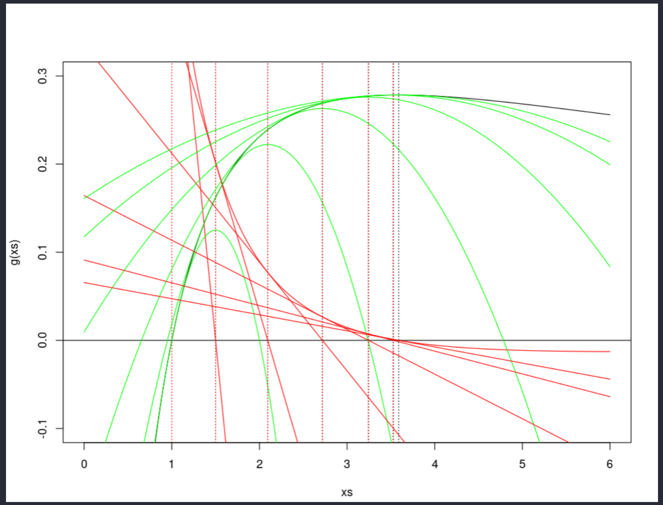




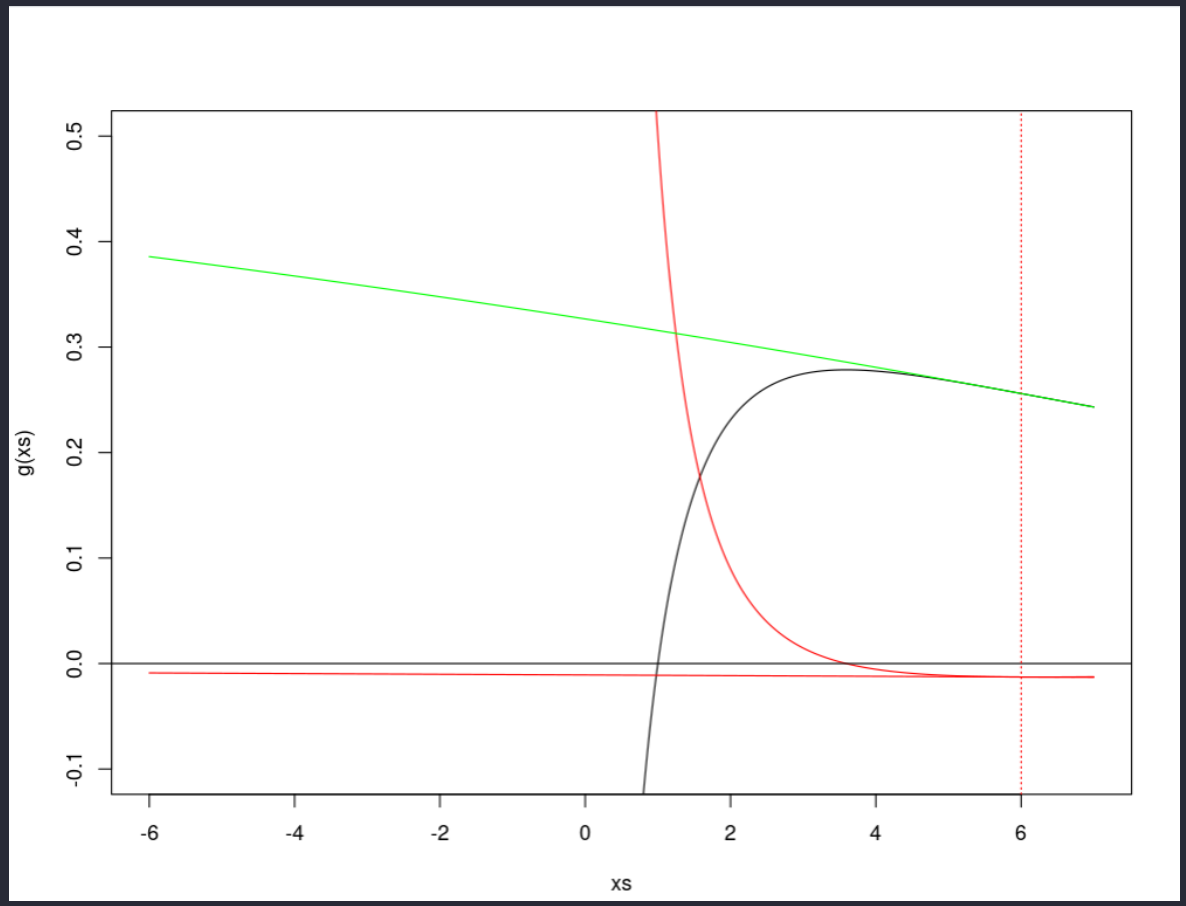








What happens if we start at $x = 6$



6 looked like a perfectly reasonable starting value.

Generalize to multivariate $\underset{\sim}{x}$

$$g(x_1, x_2)$$

$$g'(x_1, x_2) = \begin{pmatrix} \frac{\partial g}{\partial x_1} & \frac{\partial g}{\partial x_2} \end{pmatrix}$$

$$g''(x_1, x_2) = \begin{pmatrix} \frac{\partial^2 g}{\partial x_1^2} & \frac{\partial^2 g}{\partial x_1 \partial x_2} \\ \frac{\partial^2 g}{\partial x_1 \partial x_2} & \frac{\partial^2 g}{\partial x_2^2} \end{pmatrix}$$

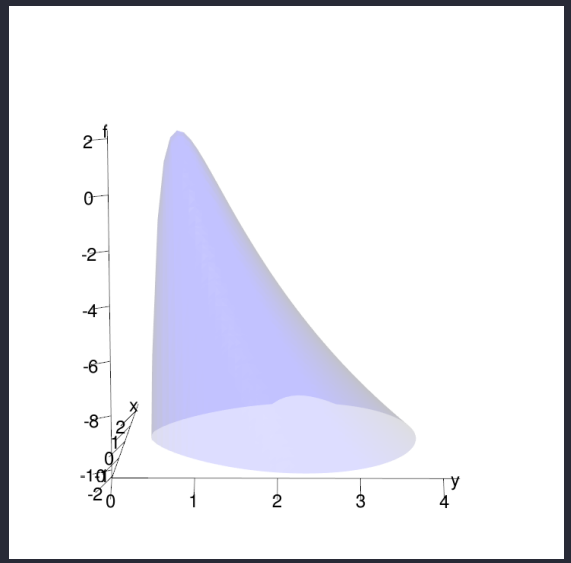
Quad approximation of $\underset{\sim}{x}_0$

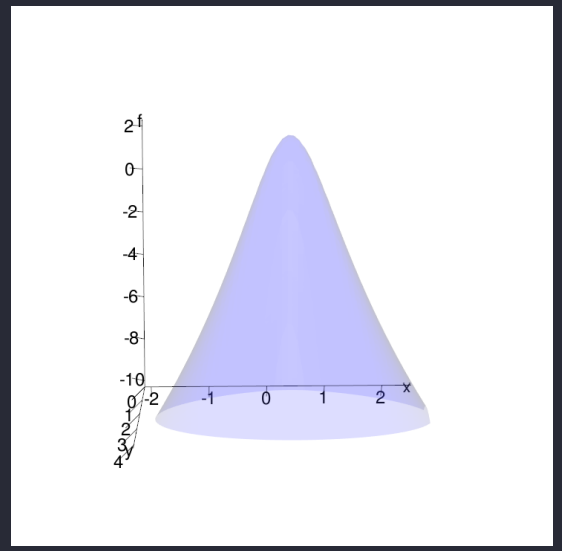
$$\hat{g}_{x_0}(\underset{\sim}{x}) = g(\underset{\sim}{x}_0) + \underbrace{g'(\underset{\sim}{x}_0)}_{1\times 2 \text{ vec}} \underbrace{(\underset{\sim}{x} - \underset{\sim}{x}_0)}_{2\times 1 \text{ vector}}$$

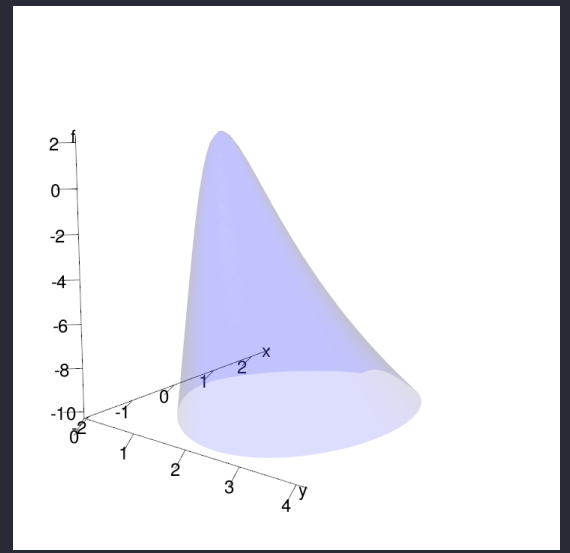
$$+ \underbrace{(\underset{\sim}{x} - \underset{\sim}{x}_0)'}_{1\times 2} \underbrace{g''(\underset{\sim}{x}_0)}_{2\times 2} \underbrace{(\underset{\sim}{x} - \underset{\sim}{x}_0)}_{2\times 1} / 2$$

Has a single max at $\hat{\underset{\sim}{x}}$ if $-g''(\underset{\sim}{x}_0)$ is positive definite

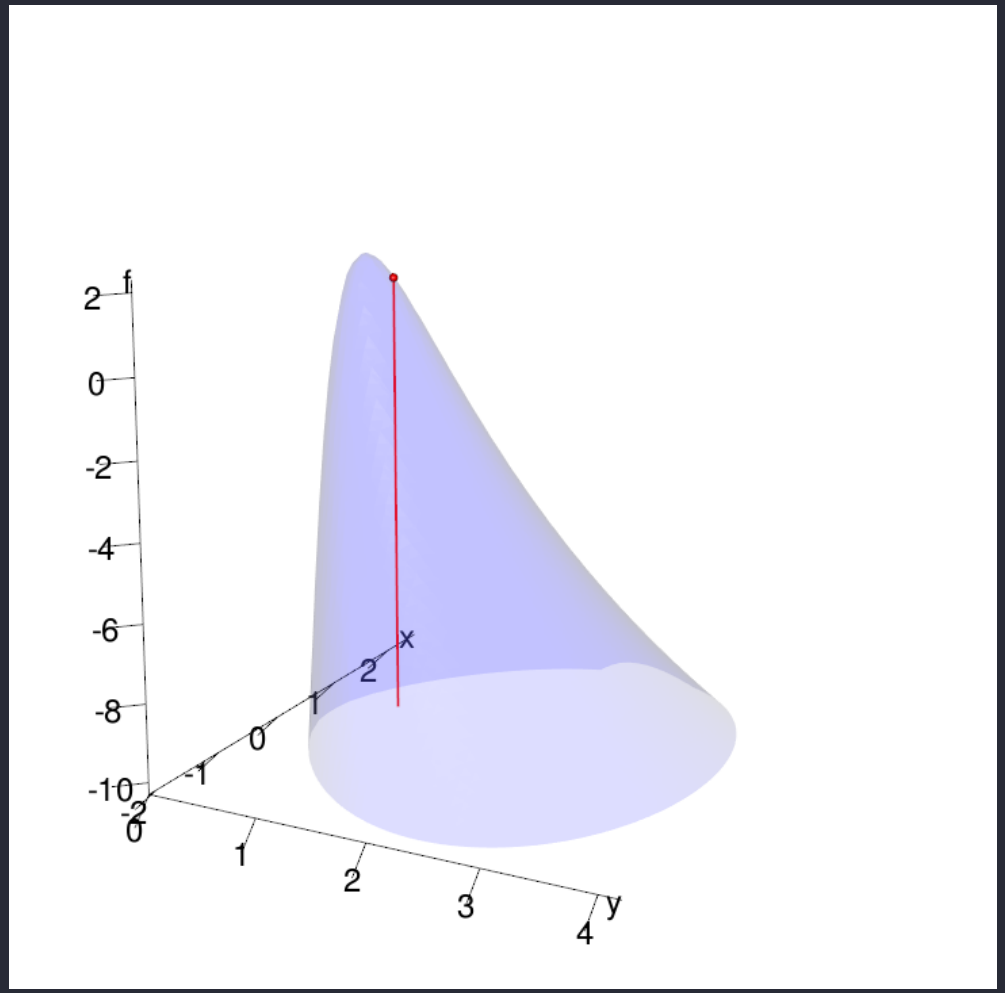
$$\hat{\underset{\sim}{x}} = \underset{\sim}{x}_0 - [g''(\underset{\sim}{x}_0)]^{-1} [g'(x_0)]^T$$

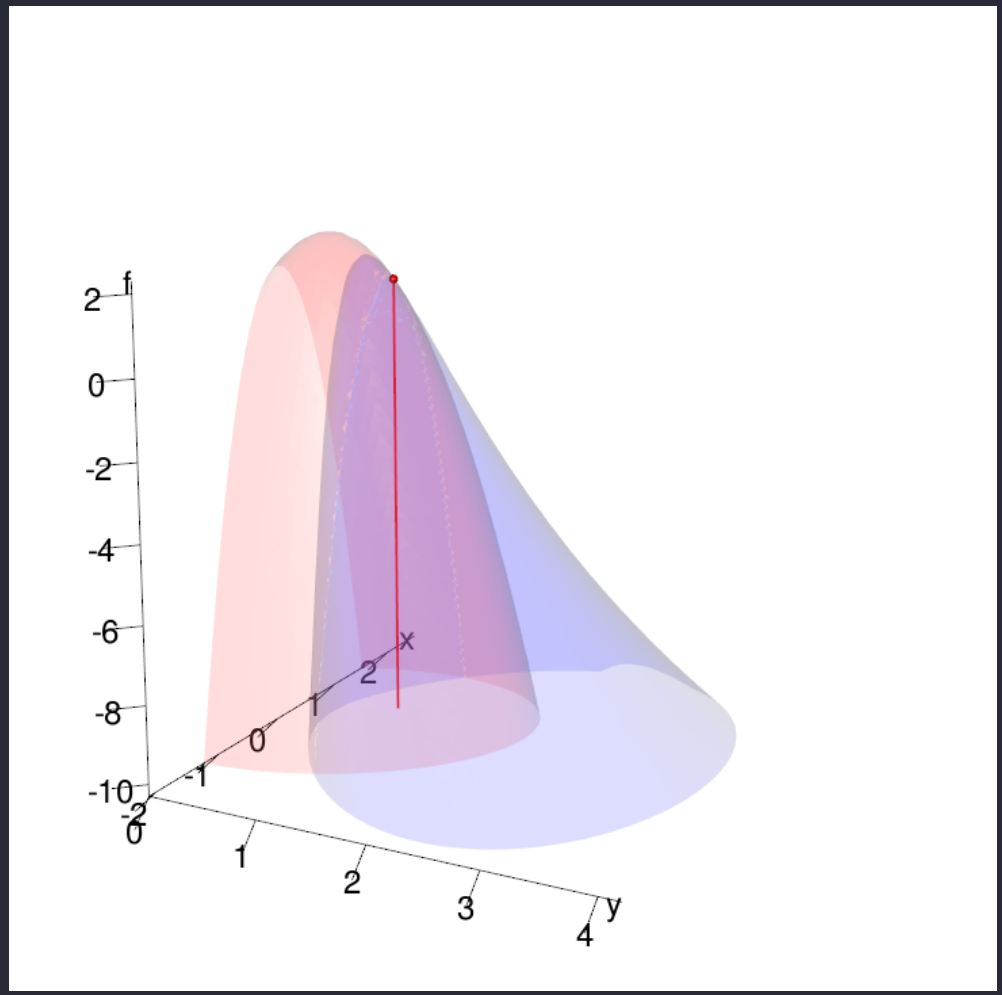

f
2
0
-2
-4
-6
-8
-10
x
2
1
0
-1
-2
0
1
2
3
4
y


f
2
0
-2
-4
-6
-8
-10
-2
-1
0
1
2
x
0
1
2
3
4
y

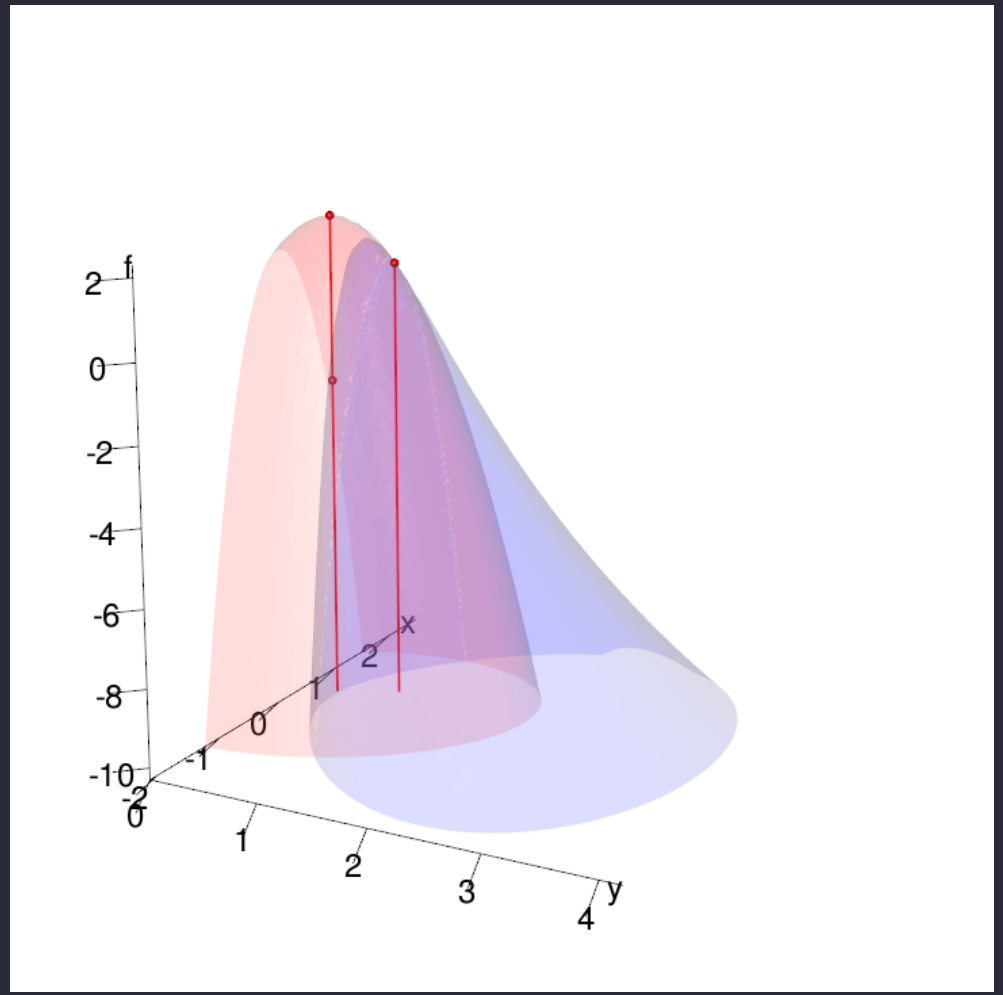

f
2
0
-2
-4
-6
-8
-10
-1
0
1
2
x
0
1
2
3
4
y

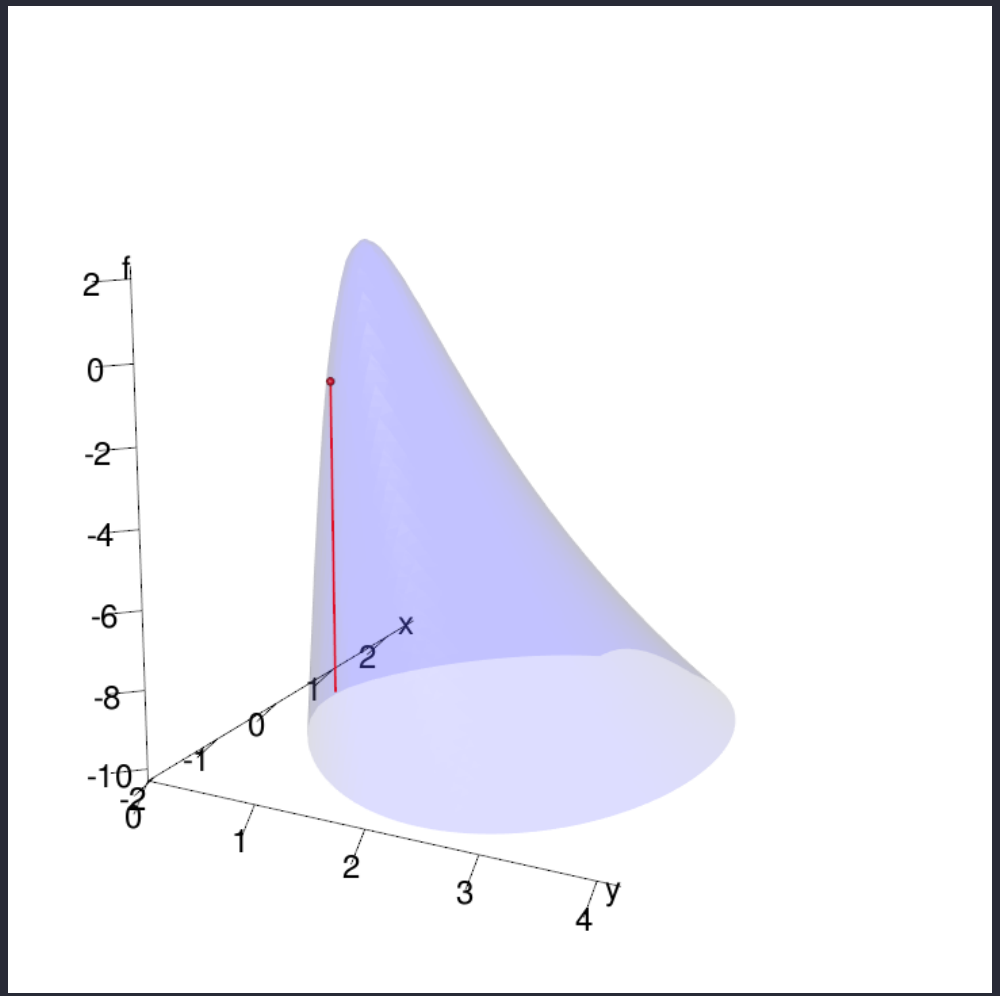
Start at $\mu_1 = .5, \sigma_1 = .9$

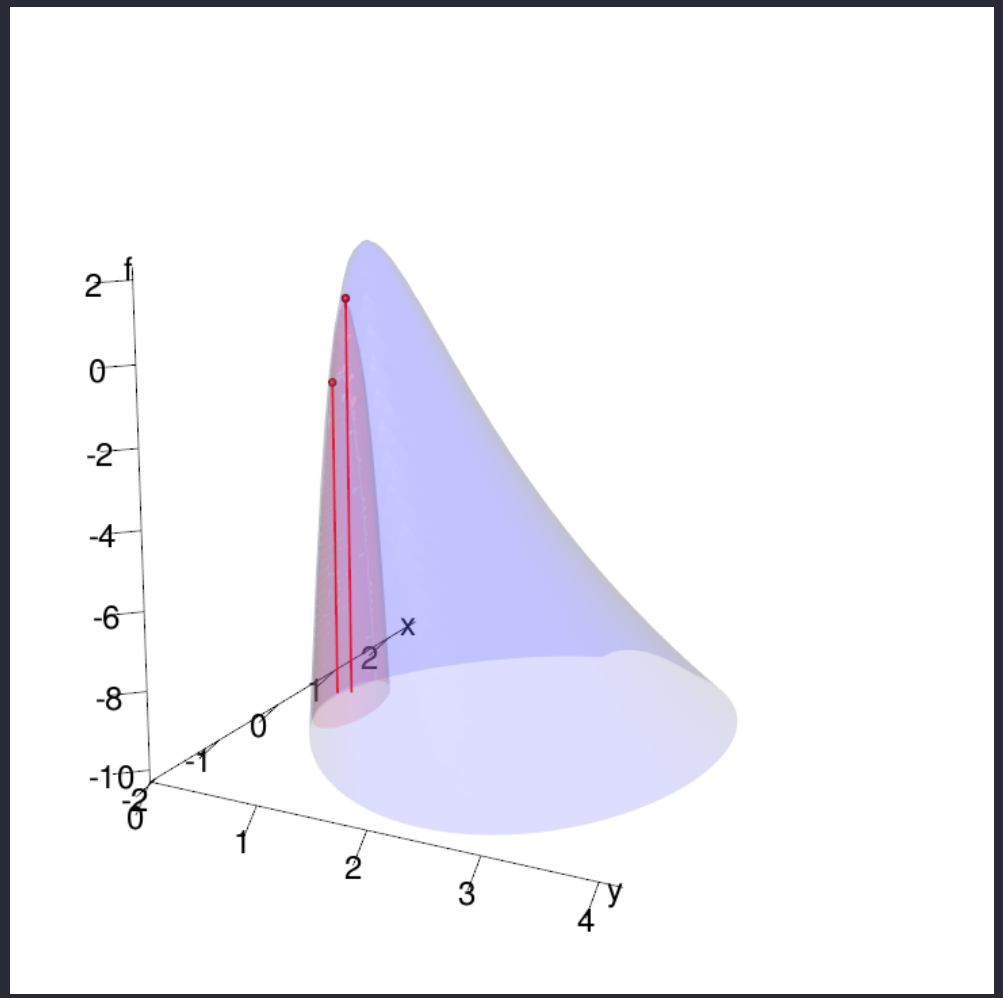


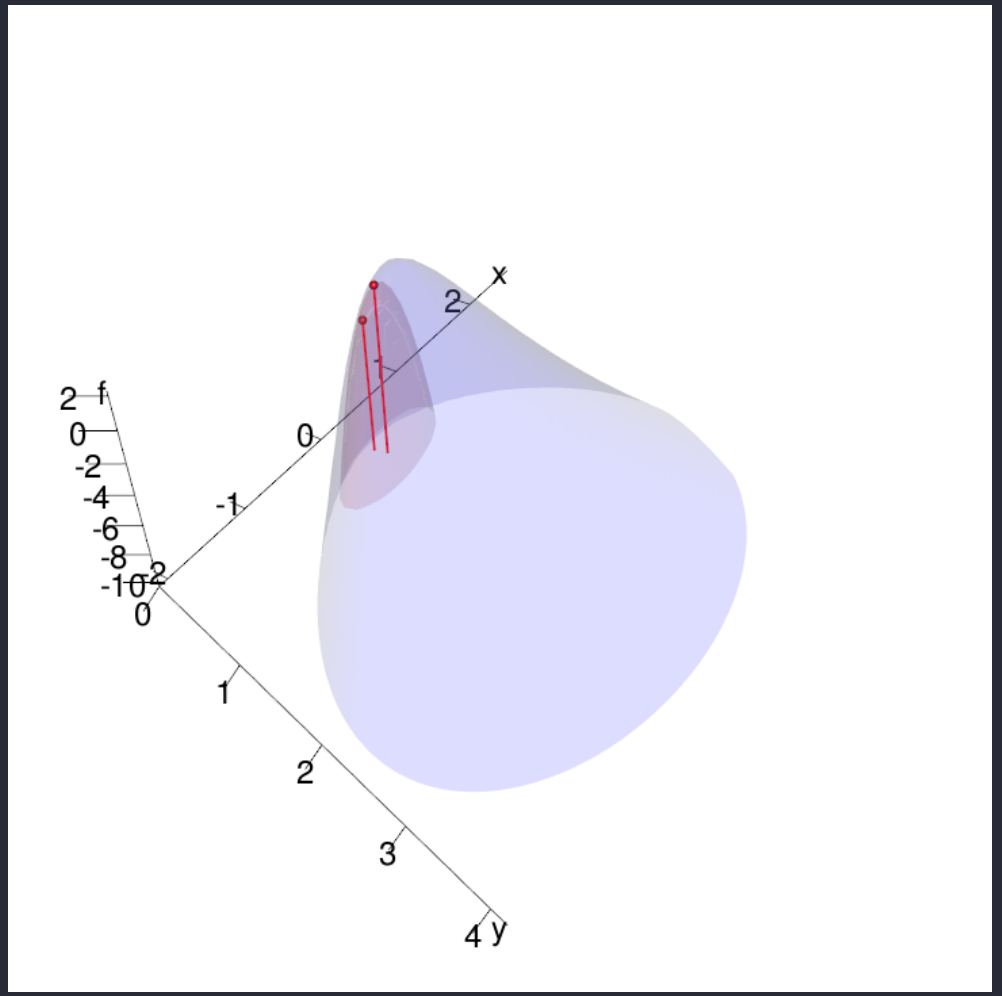

f
2
0
-2
-4
-6
-8
-10
-2
-1
0
1
2
x
0
1
2
3
4
y

```
>   Deriv(ll, x)
          [,1]      [,2]
[1,] -1.520725 -4.167008
>   drop(Deriv(function(x) Deriv(ll,x), x))    # drops dimensions of size 1
           [,1]       [,2]
[1,] -12.345679   3.379392
[2,]   3.379392 -10.801332
>   xhat <- x - c( solve(drop(Deriv(function(x) Deriv(ll,x),x))) %*% t(Deriv(ll,x)) )
>   xhat
[1] 0.2497914 0.4359312
```


f
2
0
-2
-4
-6
-8
-10
-2
-1
0
1
2
x
0
1
2
3
4
y

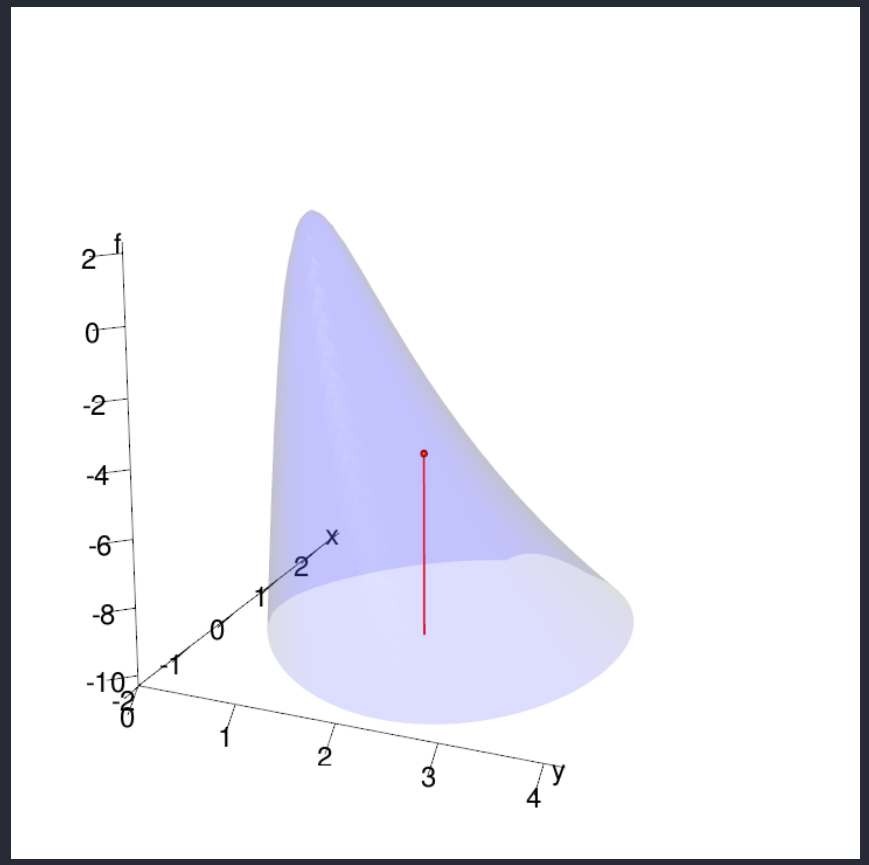

f
2
0
-2
-4
-6
-8
-10
x
-2
-1
0
1
2
y
0
1
2
3
4

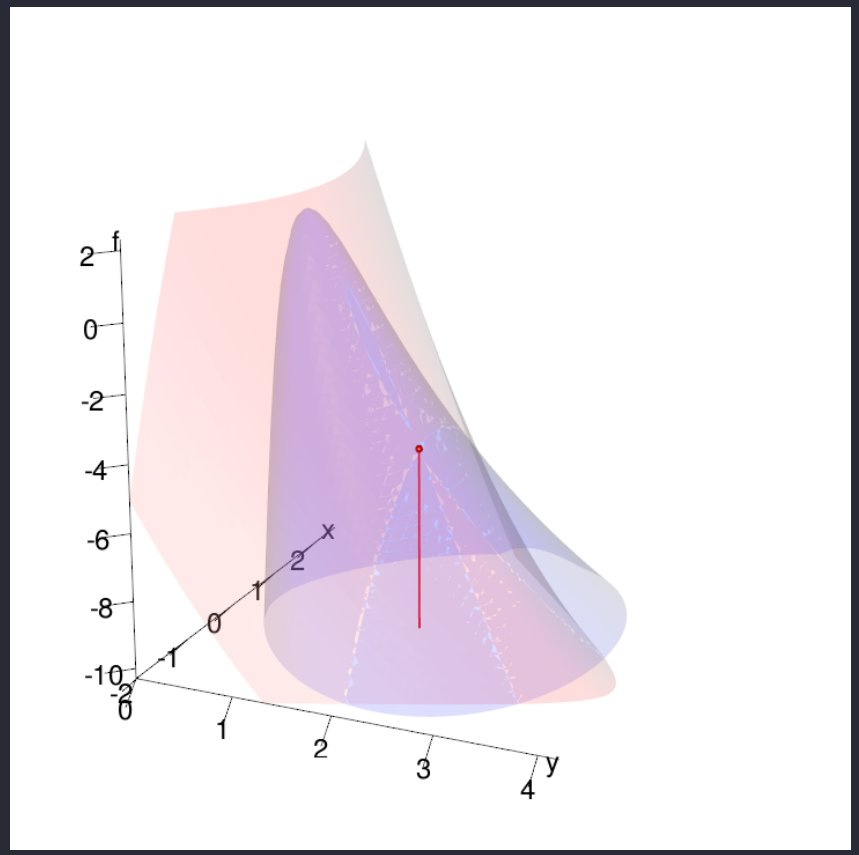




. . . gets closer to argmax(l) . . .

What if we start at $\mu_1 = 0$, $\sigma_1 = 2$






```
>   xhat <- x - c( solve(drop(Deriv(function(x) Deriv(ll,x),x))) %*% t(Deriv(ll,x)) )
>   xhat
[1] -0.6541702  4.7360095
```

. . . ? . . . ?

Newton's method needs a good starting point — close enough to the maximum for the quadratic approximation to

1) have a negative definite Hessian
   positive — for a minimum

2) have a maximum that doesn't take you outside the domain in which further quadratic approximation eventually take you to the maximum.

When will NM work?

Recall Taylor Series expansion

Not approximation

If f, f', f'' are continuous in $(x_0 \pm \varepsilon)$

$$\forall\, x \in (x_0 \pm \varepsilon),\ \exists\, x_1 \in (x_0 \pm \varepsilon)$$

$$\ni f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_1)(x - x_0)^2/2$$

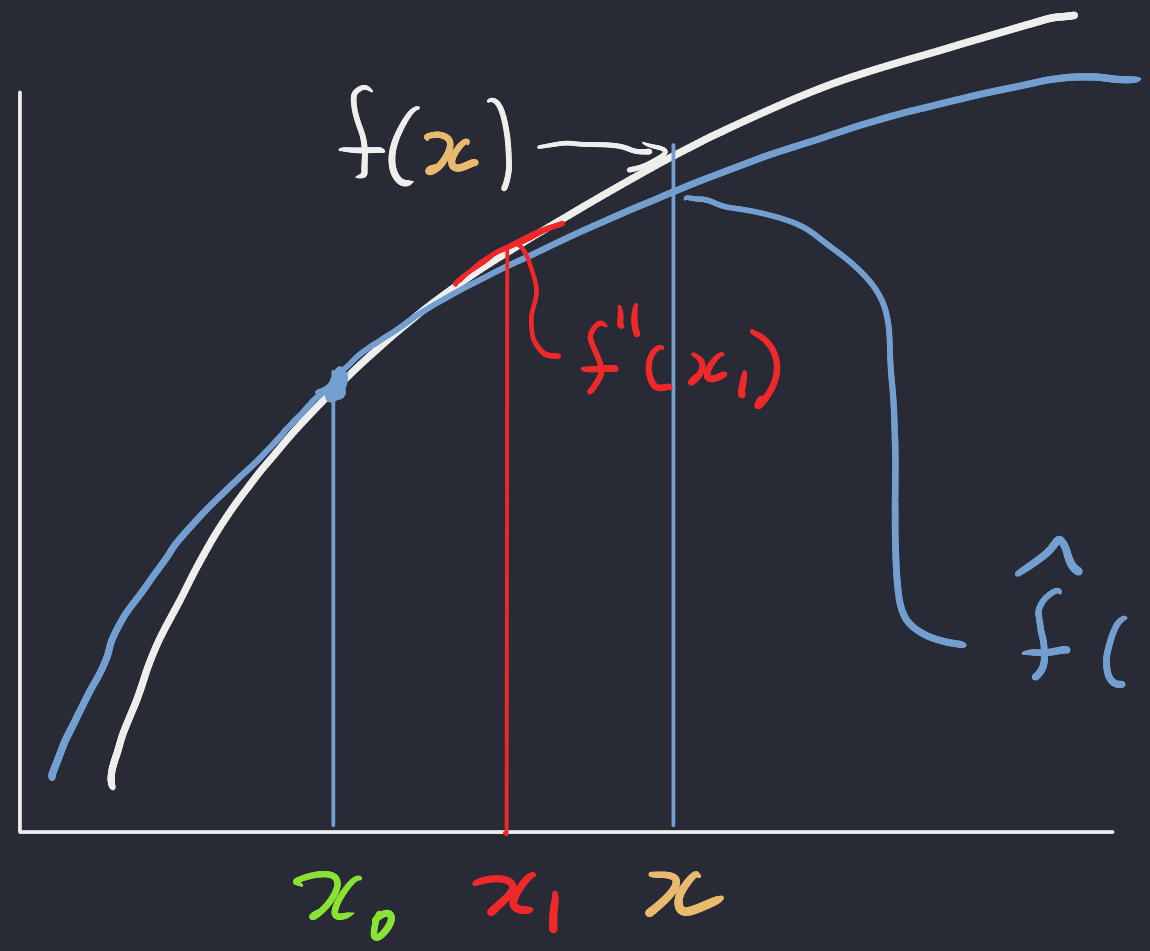
Taylor series expansion

Note:

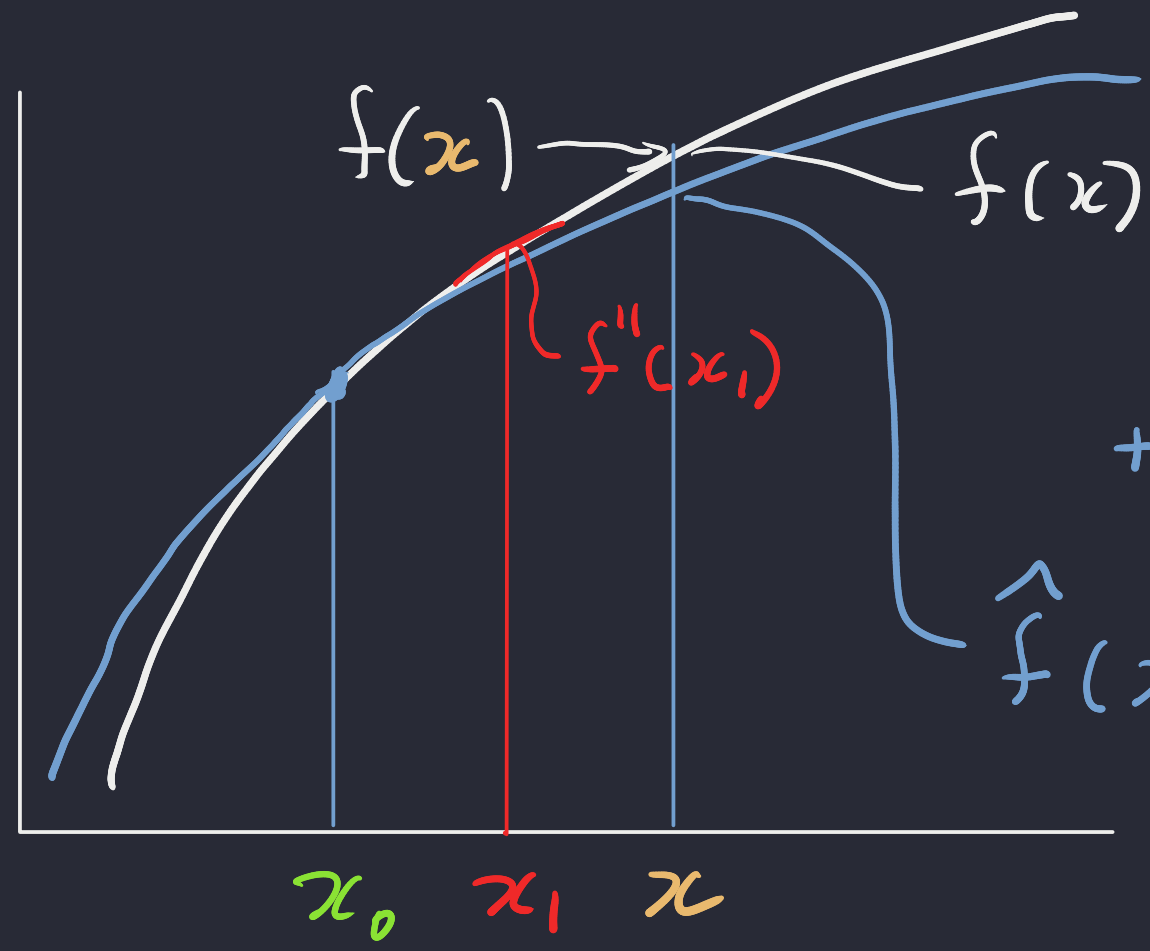
$$\hat{f}(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0)^2/2$$

← compare

Taylor series approximation



$$\hat{f}(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0)^2/2$$



$$f(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_1)(x - x_0)^2/2$$

$$\hat{f}(x) = f(x_0) + f'(x_0)(x - x_0) + f''(x_0)(x - x_0)^2/2$$

BACK TO:

When can we be sure that Newton's method will work?

Let x^* be $\operatorname{argmax}(g)$, $g \in C^3$.
So $g'(x^*) = 0$. Suppose $g''(x^*) < 0$.

By continuity, $\exists\ \varepsilon > 0$
$\ni g''(x) < 0 \quad \forall x \in (x^* - \varepsilon, x^* + \varepsilon)$

Let $\varepsilon_t = x_t - x^*$ for $|\varepsilon_t| < \varepsilon$

Apply Taylor expansion to g'

$$0 = g'(x^*) = g'(x_t) + g''(x_t)(x^* - x_t) + \frac{1}{2}(x^* - x_t)^2 g'''(x^b)$$

for x^b between x^* & x_t

Now $x_{t+1} = x_t - [g''(x_t)]^{-1} g'(x_t)$

$$(x_{t+1} - x^*) = (x_t - x^*)^2 \frac{g'''(x^b)}{2 g''(x_t)}$$

So $\varepsilon_{t+1} = \varepsilon_t^2 \dfrac{g'''(x^b)}{2g''(x_t)}$

Now, $\left|\dfrac{g'''(x^b)}{2g''(x_t)}\right|$ is continuous for $(x^b, x_t) \in \underbrace{(x^* \pm \varepsilon) \times (x^* \pm \varepsilon)}_{\text{Cartesian product}}$

So continuous on compact $S = [x^* \pm \delta] \times [x^* \pm \delta]$
if $\delta < \varepsilon \Rightarrow$ has a maximum on S, say M.

$\therefore\ |\varepsilon_{t+1}| \leq \varepsilon_t^2 M$ if $|\varepsilon_t| < \delta$

and $|\varepsilon_{t+1}| \leq \frac{1}{2}|\varepsilon_t|$ if $|\varepsilon_t| < \min\{\delta, \frac{1}{2M}\}$

Therefore $|\varepsilon_t| \to 0$ as $t \to \infty$

and $x_t \to x^*$

if we start within $\min\{\delta, \frac{1}{2M}\}$ of x^*.

When can we be sure that Newton's Method will work?

NMC 1:

Suppose: 1) $x_* \ni g'(x_*) = 0$ (such that) — { x_* is a root of g'

2) $g' \in C^2$ near x_* — { C^2: twice continuously differentiable

3) $g''(x_*) \neq 0$

Then $\exists\ \varepsilon > 0$ (there exists) — not the same as the one we started with above

$\ni$ starting from any $x_0 \in (x_* \pm \varepsilon)$ — my notation for an interval

Newton's Method will converge to x_* (such that)

Proof: above.

When can we be sure that Newton's Method will work?

NMC 1:

Suppose: 1) $\exists\ x_*$ such that $g'(x_*) = 0$ { x_* is a root of g'

2) $g' \in C^2$ near x_* { C^2: twice continuously differentiable

3) $g''(x_*) \neq 0$

Then $\exists\ \varepsilon > 0$ such that starting from any $x_0 \in (x_* \pm \varepsilon)$ (my notation for an interval)

Newton's Method will converge to x_*

Proof: Above.

NMC2: If $g \in C^3$, $g''(x_*) < 0$
and g concave down
then NM converges from anywhere.

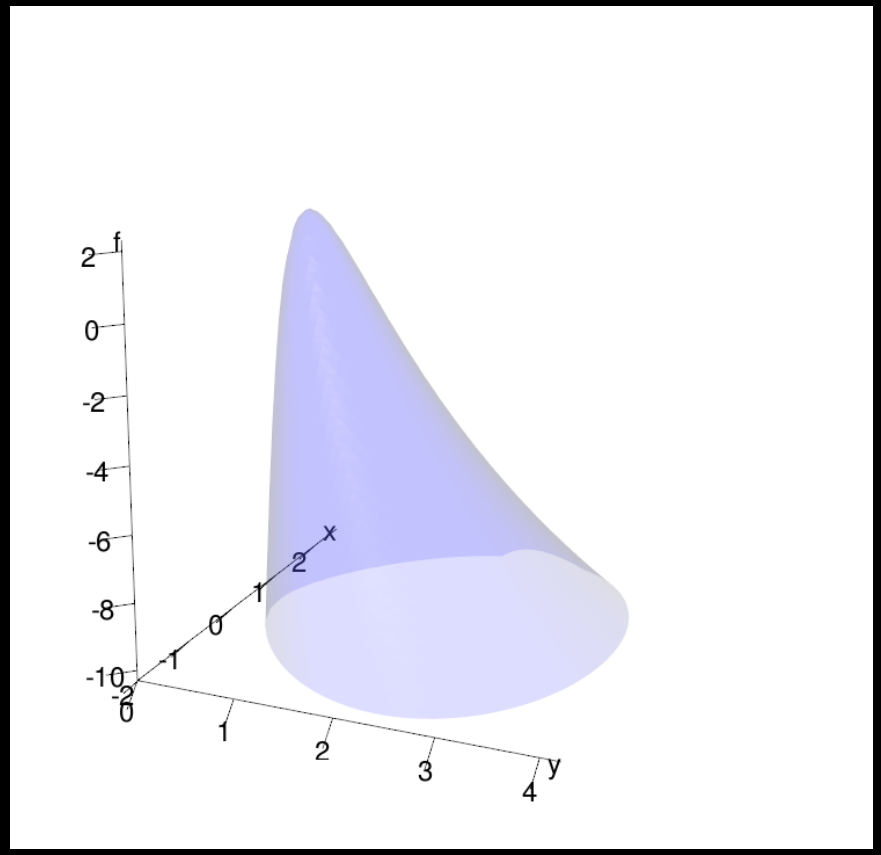
So alternative strategy:

Reparametrize g to make it concave down

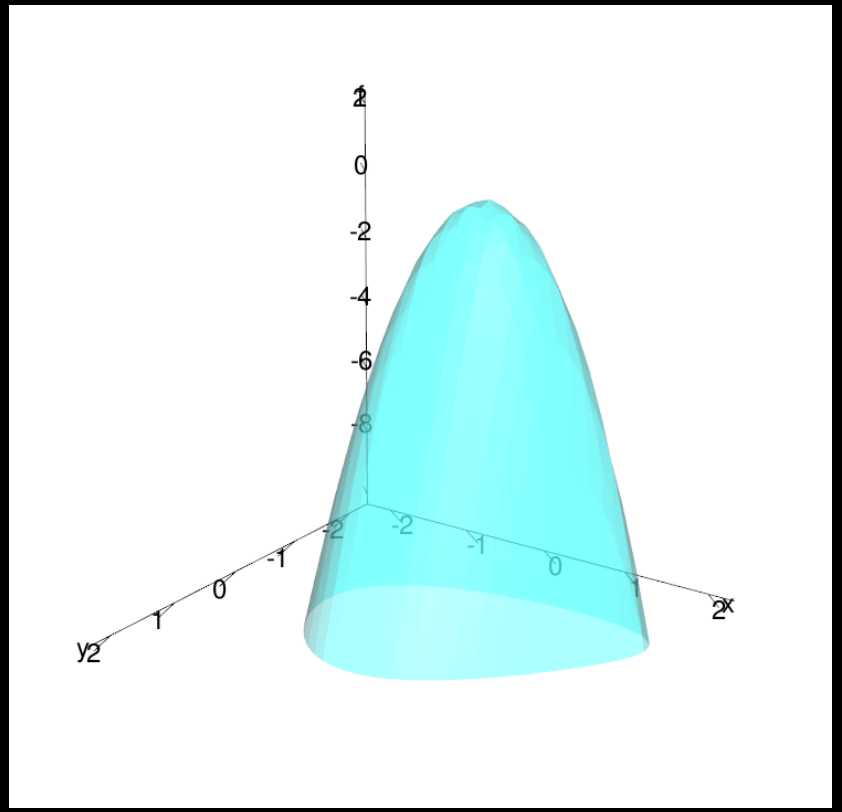
e.g. Instead of $f(x \mid \mu, \sigma)$

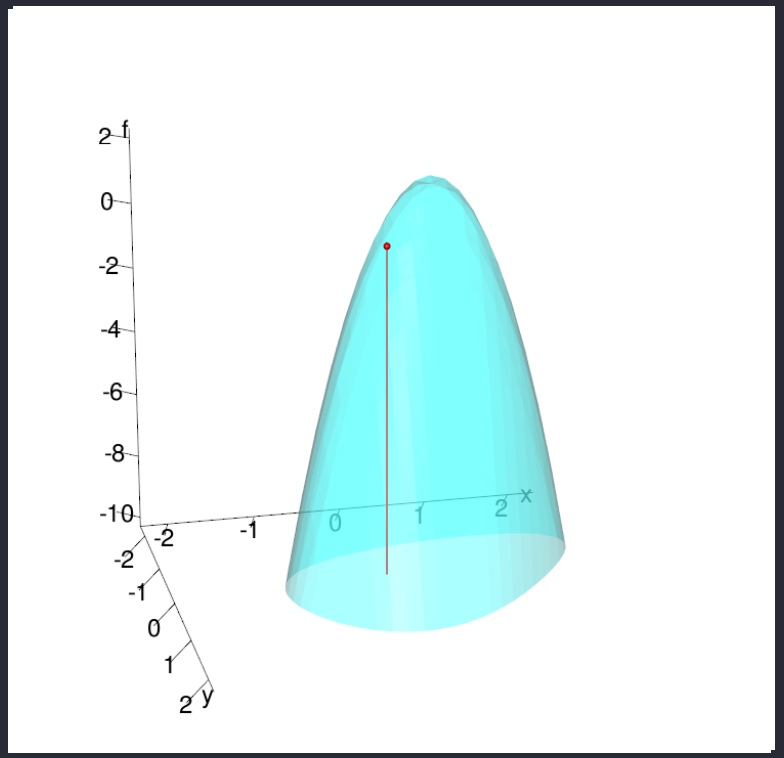
use $f(x \mid \theta, \psi)$ where $\theta = \mu/\sigma$ $\psi = \log \sigma$

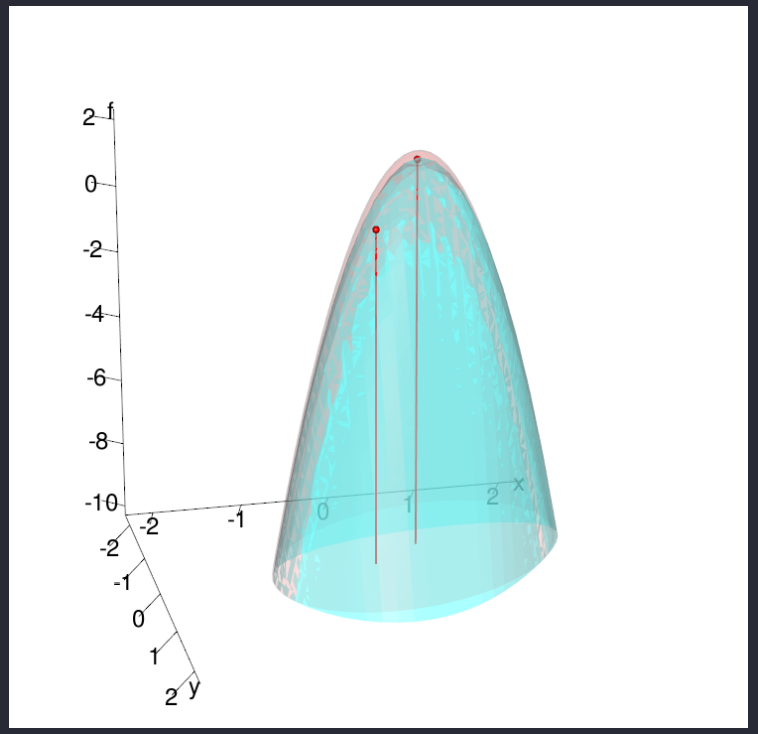
Using μ, σ:



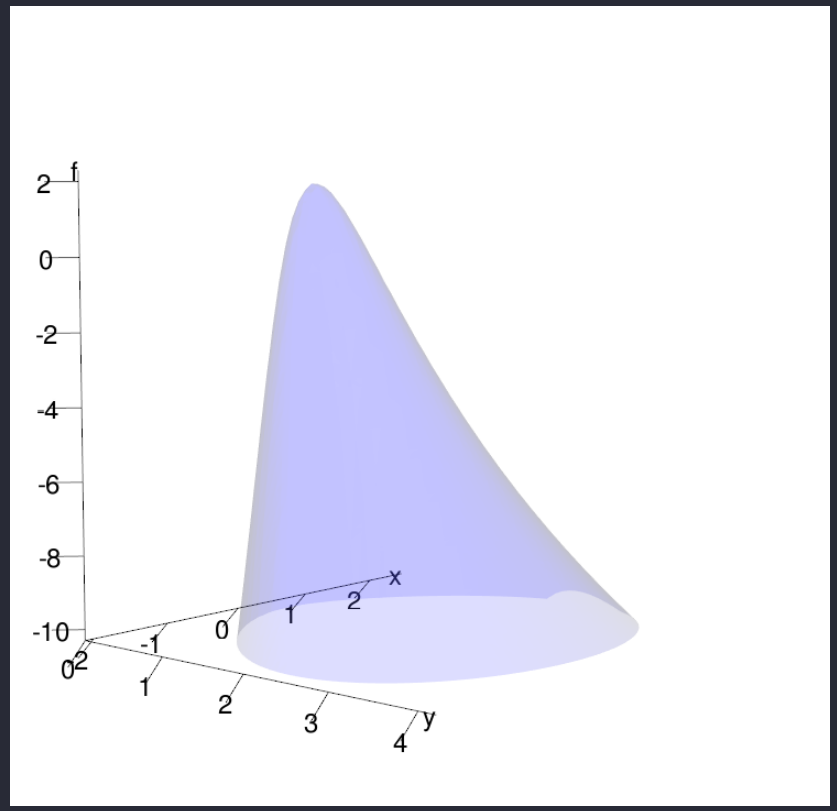
Using θ, ψ:

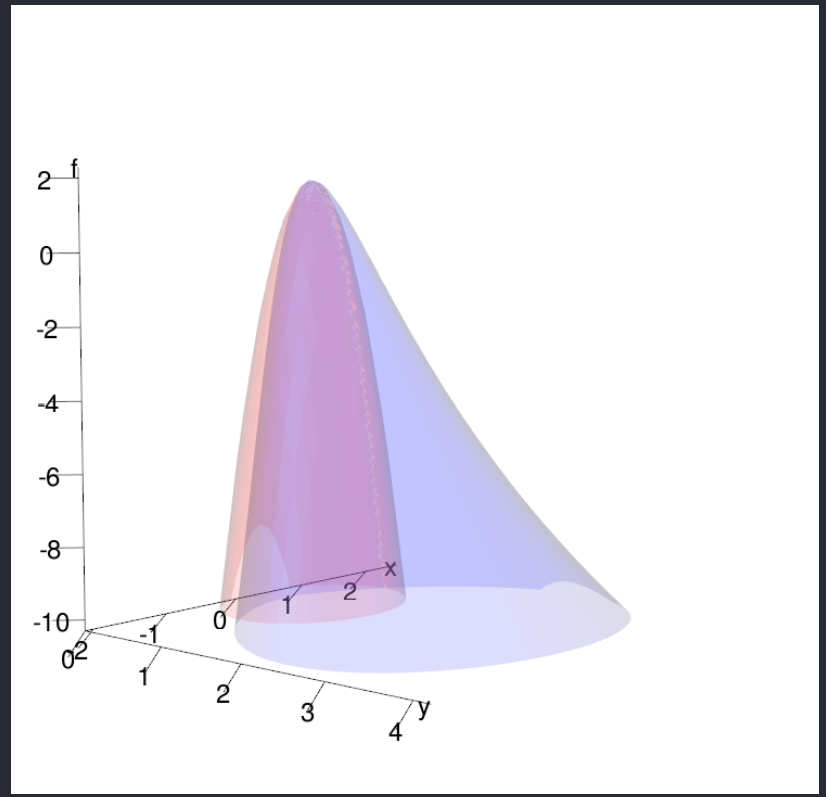


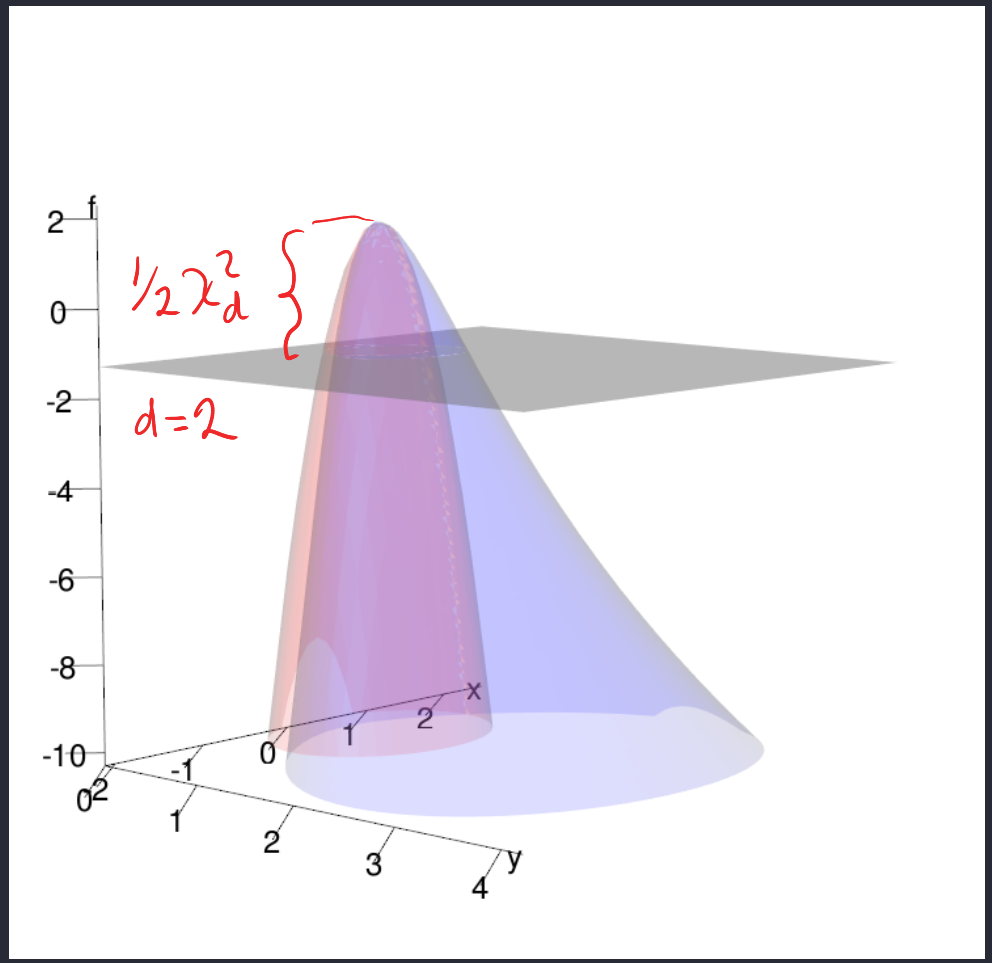

f
2
0
-2
-4
-6
-8
-10
x
-2
-1
0
1
2
y
-2
-1
0
1
2


f
2
0
-2
-4
-6
-8
-10
x
-2
-1
0
1
2
y
-2
-1
0
1
2

LRT vs. Wald Confidence regions







LR 95% CR

Wald 95% CR

- LR CR is equivariant under reparametrization.
- Wald CR is NOT but is much closer to LR CR if likelihood is reparametrized to be more quadratic.