On the numerical inversion of cumulative distribution functions

Javier Segura

Departamento de Matemáticas, Estadística y Computación.
Universidad de Cantabria, Spain

1. **Introduction and definitions**

2. **Algorithms based on initial value estimations**

3. **Iterative methods with improved global properties**
   - Halley's method
   - The Schwarzian-Newton method
   - Geometrical interpretation of the SNM
   - Applications

Introduction

Cumulative distribution functions

$$F(x) = \int_{\alpha}^{x} f(t)dt,\ f(t) > 0,\ F(\beta) = 1.$$

$F(x)$: cumulative distribution function (CDF).
$f(x)$: probability density function (PDF).

The CDF gives the probability that a random variable X with PDF f will be found to have a value less than or equal to x

Given $0 < p < 1$, the inverse of $F(x) = p$ with respect to x is an important funtion in statistics (quantile function).

An example of application is random number generation:

If U is a random uniform variable in $[0, 1]$ then $Y = F^{-1}(U)$ is a random variable distributed according to probability density function $f(t)$

A simple and well studied example is the gaussian distribution

$$F(x) = \frac{1}{\sigma\sqrt{\pi}} \int_{-\infty}^{x} e^{-\frac{(t-\mu)^2}{2\sigma^2}} dt = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right)\right]$$

where erf is the error function

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-t^2} dt$$

e$rf(x)$ is a "nearly elementary" function for which efficient methods of computation and inversion exist.

Many other CDFs can be expressed, in some limit, in terms of error functions (at least asymptotically).

The error function is a particular case of the central gamma distribution.
$P_{1/2}(x) = \text{erfc}(\sqrt{x}) = 1 - \text{erf}(\sqrt{x})$.

Cumulative γ and β distributions

The cumulative central gamma distribution is given by

$$P_a(x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt,\ x \geq 0.$$

The cumulative central beta distribution is defined as

$$B_{a,b}(x) = \frac{1}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} dt,\ B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},\ x \in [0,1]$$

The cumulative noncentral γ distribution can be written

$$P_a(x,y) = y^{\frac{1}{2}(1-a)} \int_0^x t^{\frac{1}{2}(a-1)} e^{-t-y} I_{a-1}\left(2\sqrt{yt}\right)\ dt, x \geq 0$$

The cumulative noncentral beta distribution can be written as

$$B_{a,b}(x,y) = \frac{e^{-y/2}}{B(a,b)} \int_0^x t^{a-1}(1-t)^{b-1} M\left(a+b, a, \frac{yt}{2}\right) dt,\ x \in [0,1]$$

The problem is to invert these CDFs and to do this as efficiently as possible.

Of course, the problem of computing the functions comes first and without algorithms for an accurate computation of $P_a(x, y)$ and $B_{a,b}(x, y)$ it is not possible to build efficient algorithms for their inversion. Precisely the main cost in the inversion algorithms is given by the computation of the functions. Derivatives are cheaper.

An analysis of the situation revealed that:

1. Algorithms existed both for the computation and inversion off the CDF for the central case, but asymptotic approximations can improve both algorithms.
2. Secant, Newton and some third order methods were used for inversion starting with some initial estimations (not always accurate). Convergence was checked experimentally.
3. No reliable algorithms were available for the computation of the CDFs for the noncentral case and the available software had a limited range of validity (and also some inaccuracies).
4. Algorithms for the inversion (with respect to both x and y) in the noncentral case also had a limited range of validity and no published software was available.

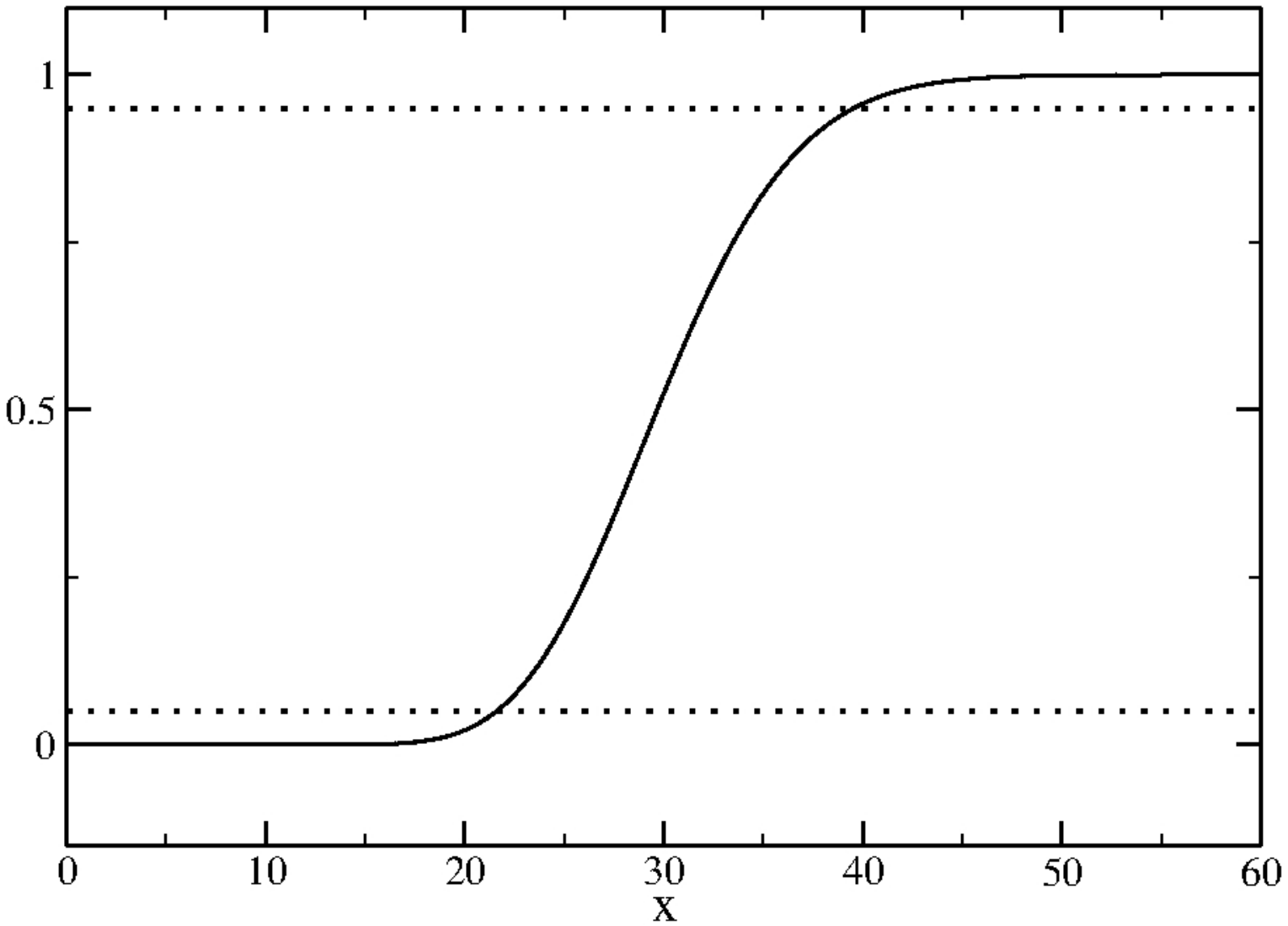
Common characteristics of these (and other CDFs):

- Unimodal: the PDF has a single extremum, and it is a maximum; therefore The CDF has only one inflection point.

- The graph of the CDF has flat ends (tails).

We consider as example the case of the central gamma distribution.

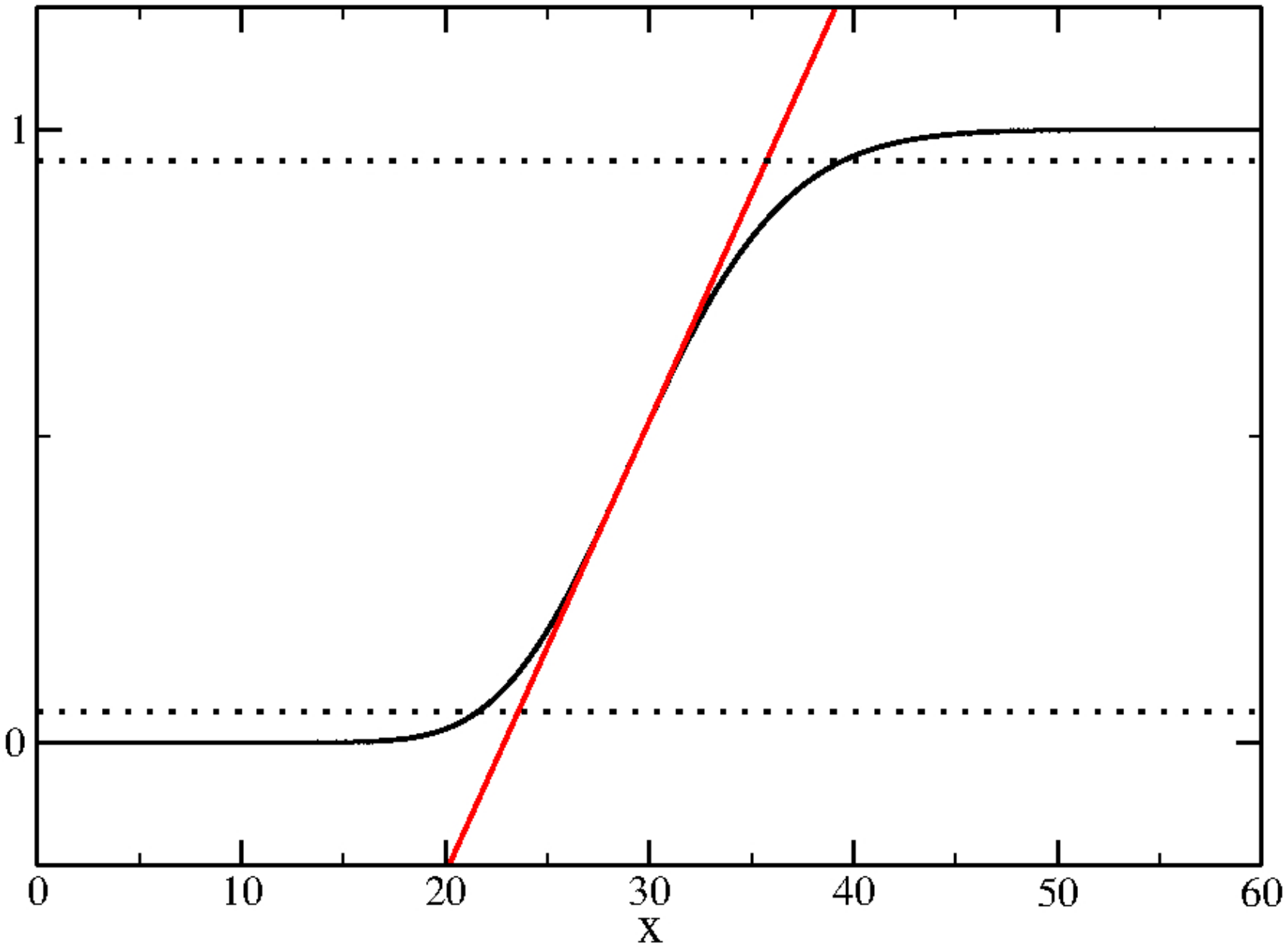
Graph of $P_a(x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$, $a = 30$



$$P_a(x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$$

The PDF has its maximum at $x = a - 1$ when $a > 1$, where the CDF has its inflection point.
Therefore:

1. The inversion of $P_a(x) = p$ by Newton method converges monotonically to the solution if the starting value x_0 is closer to $a - 1$ than the root.
2. With $x_0 = a - 1$ we have convergence for any value of $p \in (0, 1)$ and $a > 1$.


1
0
0
10
20
30
40
50
60
x

$$P_a(x) = \frac{1}{\Gamma(a)} \int_0^x t^{a-1} e^{-t} dt$$

The PDF has its maximum at $x = a - 1$ when $a > 1$, where the CDF has its inflection point.
Therefore:

1. The inversion of $P_a(x) = p$ by Newton method converges monotonically to the solution if the starting value x_0 is closer to $a - 1$ than the root.
2. With $x_0 = a - 1$ we have convergence for any value of $p \in (0, 1)$ and $a > 1$. But this is slow, particularly for the tails.

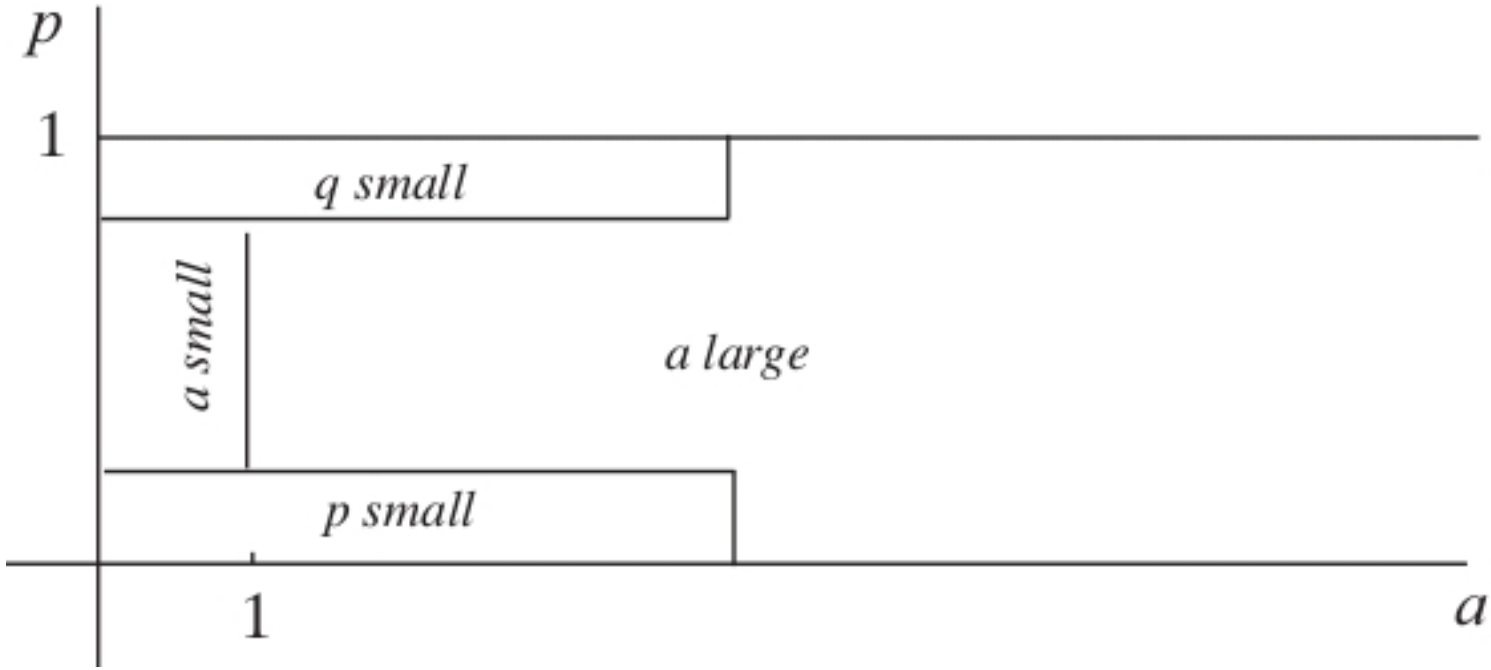
More accurate starting values are convenient. And because the derivatives are cheap to compute an efficient algorithm could use:

1. Sufficiently accurate analytical starting values
2. High order Newton-like methods.

Algorithms based on initial value estimations

The computation of initial estimations has different difficulty depending on the distribution to invert and very specifically on the number of parameters

For the inversion of the central gamma distribution w.r.t. x ($P(a, x) = p = 1 - q$), different type of starting values are chosen according to the next schematic figure.



Next we give two examples (small p and large a).

For small p (therefore small x), we can try to invert approximatelly

$$P_a(x) = \frac{x^a}{\Gamma(a)} \sum_{n=0}^{\infty} \frac{(-1)^n x^n}{(a+n)n!} = p \tag{1}$$

and then

$$x = r\left(1 + \sum_{n=1}^{\infty} \frac{a(-1)^n x^n}{(a+n)n!}\right)^{-1/a}, \quad r = (p\Gamma(1+a))^{1/a}, \tag{2}$$

We can iterate the resulting fixed point iteration (truncating the sum).
Alternatively, we write

$$x = r + \sum_{n=2}^{\infty} c_k r^k,$$

The first few coefficients are

$$c_2 = \frac{1}{a+1}, \qquad c_3 = \frac{3a+5}{2(a+1)^2(a+2)},$$

$$c_4 = \frac{8a^2+33a+31}{3(a+1)^3(a+2)(a+3)}, \qquad c_5 = \frac{125a^4+1179a^3+3971a^2+5661a+2888}{24(1+a)^4(a+2)^2(a+3)(a+4)}.$$

Large p (small $q = 1-p$) can be solved in a similar way (but using asymptotic expansions).

The approximate inversion of $P_a(x) = p$ for large a (but not so large) starts from (Temme, 1979)

$$P_a(x) = \text{erfc}(-\eta\sqrt{a/2}) - R_a(\eta), \frac{1}{2}\eta^2 = \lambda - 1 - \log\lambda$$

$\lambda = a/x$ and $\text{sign}(\eta) = \text{sign}(\lambda - 1)$ and $R_a(\eta)$ is an (asymptotic) series in inverse powers of a (going to 0 as $a \to +\infty$).

We treat the asymptotic series as a perturbation. A first approximation for η_0 is obtained by inverting

$$\text{erfc}(-\eta_0\sqrt{a/2}) = p$$

Then we write η (the solution of $P_a(x) = p$) as an asymptotic series

$$\eta(p, a) = \eta_0(p, a) + \epsilon(\eta_0, a), \ \epsilon(\eta_0, a) = \sum_{j\in\mathbb{N}} \epsilon_j(\eta_0, a)a^{-j}$$

with coefficients that can be explicitly obtained in terms of η_0. After the value η is computed (aproximatelly), we obtain x (again by numerical inversion).

Although efficient, this is much more involved than the cases for p or $1 - p$ small.

The final steps in the construction of the algorithm are:

1. Starting values: select carefully the different analytical inversion methods for the different values of the parameters.
2. Test the convergence of the iterative method (we used a fifth order Newton-like method for the central gamma case).
3. Find an alternative iterative method when convergence fails (rare, but it happens sometimes, for instance for very small p). Normally, this involves using lower order iterative methods.

Some references:

A. R. DiDonato, A. H. Morris. ACM Trans. Math. Software 12 (1986) 377–393. (Central gamma distribution)

A. Gil, JS, N.M. Temme. SIAM J Sci Comput 34(6) (2012) A2965-A2981. (Central gamma distribution)

A. Gil, JS, N.M. Temme. Comput. Phys. Commun. (in press) (Algorithms for central and noncentral gamma distributions)

Iterative methods with improved global properties

It would be desirable to consider methods with better non-local convergence properties than higher order Newton, particularly for the more complex cases (non-central).

Computing starting values is sometimes hard. It would be desirable that accurate starting values are not needed (at least for the harder cases).

As starting point we consider Halley's method which, as we will see, has generally good non-local properties for the problems we are considering.

Halley's method (HM) is the following fixed point method:

$$x_{n+1} = g(x_n),\ g(x) = x - f(x)\left(f'(x) - \frac{f''(x)}{2f'(x)}f(x)\right)^{-1}$$

Theorem

Let f with $f' \neq 0$ and f''' continuous in an interval J and let $\alpha \in J$ such that $f(\alpha) = 0$. Then, if $\{f, x\} < 0$ in J the HM converges monotonically to α for any starting value $x_0 \in J$.

$\{f, x\}$ is the Schwarzian derivative of f with respect to x, that is:

$$\{f, x\} = \frac{f'''}{f'} - \frac{3}{2}\left(\frac{f''}{f'}\right)^2.$$

For the case of $f(x) = P_a(x) - p$ we have $\{f, x\} = -\frac{1}{2}\left(1 + 2\frac{1-a}{x} + \frac{a^2-1}{x^2}\right)$, which is smaller that zero for $a > 1$ and with a maximum at $a + 1$. Then:

The HM converges to the solution of $P_a(x) = p$ for any $x_0 > 0$ and $p \in (0, 1)$ if $a \geq 1$.

But it is not always faster than Newton's if we don't have sufficiently accurate starting values, as we see next.

Halley's method (HM) is exact for the functions

$$h(x) = \frac{x + A}{Bx + C} \tag{3}$$

(gives the root $x = -A$ in one iteration). From this follows the geometric interpretation:

Theorem

Let $h(x)$ as in (3) and define $y(x) = h(x - x_n)$, the HM is obtained by

1. *$y(x_n) = f(x_n)$, $y'(x_n) = f'(x_n)$, $y''(x_n) = f''(x_n)$ and $y'''(x_n) = f'''(x_n)$ (thus determining the three constants)*
2. *Obtaining x_{n+1} from $y(x_{n+1}) = 0$.*

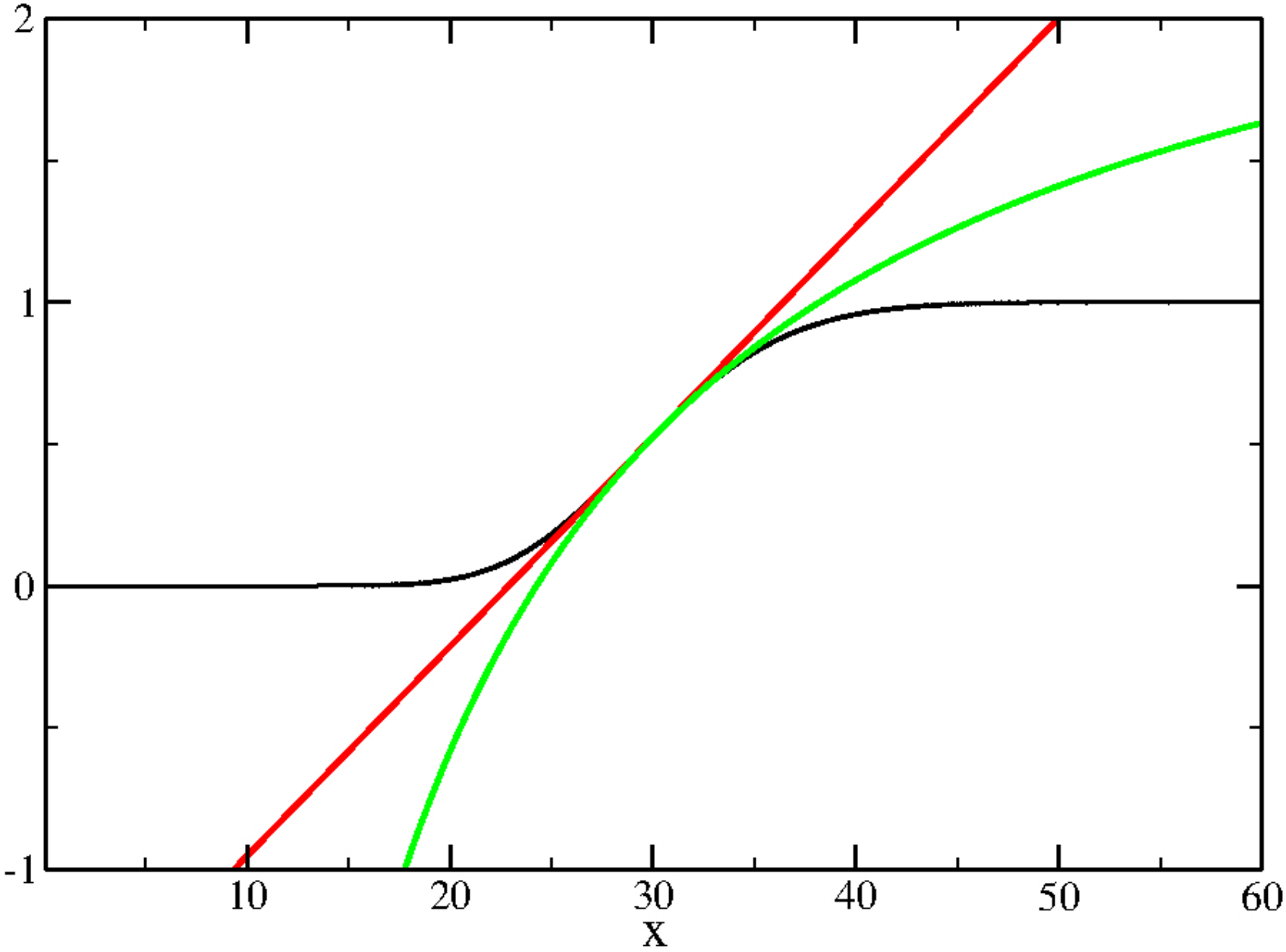
The three constants are given by

$$A = \frac{2f(x_n)f'(x_n)}{D(x_n)},\ B = -\frac{f''(x_n)}{D(x_n)},\ C = \frac{2f'(x_n)}{D(x_n)} \tag{4}$$

where

$$D(x_n) = 2f'(x_n)^2 - f(x_n)f''(x_n) \tag{5}$$

The HM is also known as the method of tangent hyperbolas


2
1
0
-1
10
20
30
40
50
60
X

The Schwarzian Newton method.

First we reconsider the HM and some of its properties.

Let f be sufficiently differentiable. Our aim is to solve $f(x) = 0$. We write

$$f''(x) + B(x)f'(x) = 0, B(x) = -f''(x)/f'(x),$$

which is a second order ODE that we can transform to normal form by setting

$$\Phi = \exp\left(\frac{1}{2}\int B\right) f = f/\sqrt{|f'|}$$

This leads to

$$\Phi'' + \Omega\Phi = 0, \ \Omega = -\frac{1}{4}B^2 - \frac{1}{2}B' = \frac{1}{2}\{f, x\}. \tag{6}$$

where $\{f, x\}$ is the Schwarzian derivative of f with respect to x.

As is well know, the application of Newton's method to the function Φ leads to the HM. And considering (6) the global convergence of HM when $\{f, x\} < 0$ is explained.

The HM is a third order method (if $\Phi(\alpha) = 0$ then $\Phi''(\alpha) = 0 \Rightarrow g'(\alpha) = g''(\alpha) = 0$).

Now, we improve HM by integrating approximatelly the Riccati equation associated to $\Phi'' + \Omega\Phi = 0$.

$$\Phi'' + \Omega\Phi = 0 \Rightarrow h'(x) = 1 + \Omega(x)h(x)^2,\ h = \Phi/\Phi'.$$

Now let α such that $h(\alpha) = 0$ (then $f(\alpha) = 0$). IF $\Omega(x) > 0$, we have:

$$x - \alpha = \int_\alpha^x \frac{h'(t)}{1 + \Omega(t)h^2(t)}dt \approx \frac{1}{\sqrt{\Omega(x)}} \arctan(\sqrt{\Omega(x)}h(x)). \tag{7}$$

where the approximation consists in taking $\Omega(x)$ constant in the integration.
This suggests:

$$g(x) = x - \frac{1}{\sqrt{\Omega}} \arctan\left(\sqrt{\Omega}\frac{\Phi}{\Phi'}\right), \tag{8}$$

where

$$\frac{\Phi}{\Phi'} = \frac{f(x)}{f'(x) - \dfrac{f''(x)}{2f'(x)}f(x)}.$$

The Schwarzian-Newton method (SNM)

$$x_{n+1} = g(x_n),\ g(x) = x - \arctan\left(\frac{1}{2}\{f, x\}, \frac{f}{f' - \frac{f''}{2f'}f}\right),$$

(JS, 2015), where

$$\operatorname{arctan}(\lambda, x) = \begin{cases} \frac{1}{\sqrt{\lambda}}\arctan(\sqrt{\lambda}x) & , \quad \lambda > 0, \\ x & , \quad \lambda = 0 \\ \frac{1}{\sqrt{-\lambda}}\operatorname{arctanh}(\sqrt{-\lambda}x) & , \quad \lambda < 0 \end{cases}$$

and we use a similar definition for $\tan(\lambda, x)$

The method is exact for functions with constant Schwarzian derivative. The functions with constant Schwarzian derivative are

$$h(x) = \frac{\tan(\lambda, x) + A}{B\tan(\lambda, x) + C},$$

with $\{h, x\} = 2\lambda$.
As a particular case, the functions with zero Schwarzian derivative are

$$h(x) = \frac{x + A}{Bx + C},$$

and the HM is exact for these functions (as well as the SNM, which coincides in this case).

The SNM has a geometrical interpretation too.

Theorem

Let

$$h(x) = \frac{\tan(\lambda, x) + A}{B\tan(\lambda, x) + C},\ y(x) = h(x - x_n),$$

the SNM is obtained by

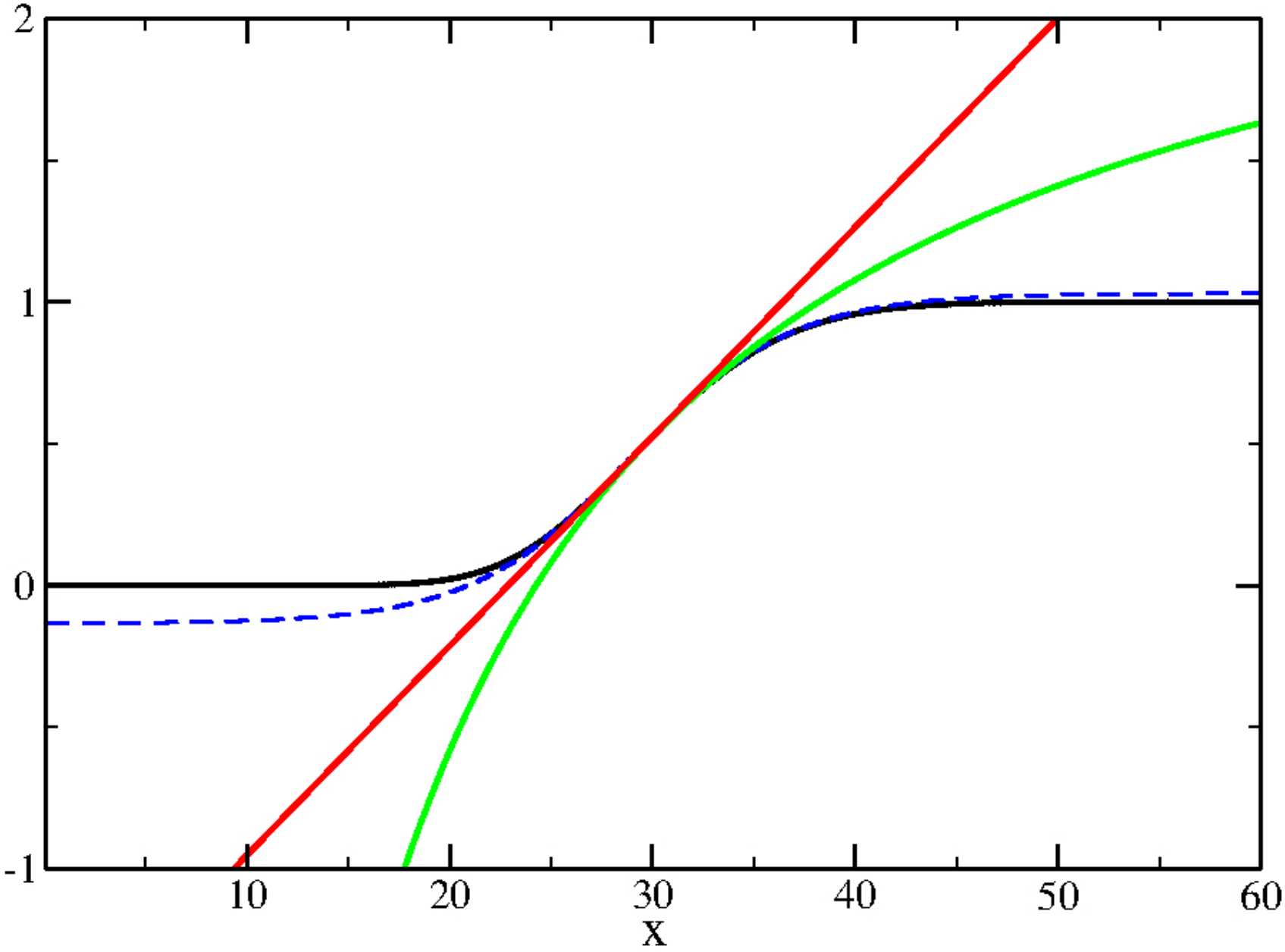
1. *Setting* $y(x_n) = f(x_n)$, $y'(x_n) = f'(x_n)$, $y''(x_n) = f''(x_n)$ *and* $y'''(x_n) = f'''(x_n)$ *(thus determining the four constants)*
2. *Obtaining* x_{n+1} *from* $y(x_{n+1}) = 0$

The constant λ *is given by*

$$\lambda = \Omega(x_n),\ \Omega(x) = \frac{1}{2}\{f, x\} \tag{9}$$

and the other three constants A, B and C have the same expression as for the HM.

Let us compare the osculating curves at $x = a + 1$ for the Newton method, the HM and the SNM.


2
1
0
-1
10
20
30
40
50
60
X

The osculating curve for the SNM at $x = a + 1$ is much closer to the graph of $P_a(x)$, particularly at the tails. In addition, the SNM has good non-local properties if $\{f, x\}$ satisfies some monotonicity properties.

Theorem

Let f such that $f' \neq 0$ and f''' continuous in an interval J and let $\alpha \in J$ such that $f(\alpha = 0)$, then if $\{f, x\}$ is decreasing in $I = [a, \alpha] \subset J$ and $\{f, x\} < 0$ in J the SNM converges monotonically to α for any starting value $x_0 \in [a, \alpha]$. If $\{f, x\} > 0$ in part of the interval, the same is true if, in addition, the SNM iteration satisfies $g(a) > a$.

(and a similar result can be given for $\{f, x\}$ increasing).

Corollary

If $\{f, x\}$ has one and only one extremum at $x_e \in J$ and it is a maximum, then

1. *If $\{f, x\}$ is negative the SNM converges monotonically to α starting from $x_0 = x_e$.*
2. *If $(x_e - \alpha)(x_e - g(x_e)) > 0$ the SNM converges monotonically to α starting from $x_0 = x_e$.*

- For $f(x) = P_a(x) - p$, $\{f, x\}$ is negative when $a > 1$ and with a maximum at $x = a + 1$. Therefore, the starting value $x_0 = a + 1$ ensures monotonic convergence.
- The convergence is fast if p is not too small or large and always better than NM or HM. If $0.05 < p < 0.95$ three iterations starting with $x_0 = a + 1$ are enough for 15-16 digits accuracy.
- For $0 < a < 1$ the change $z(x) = \log(x)$ enables monotonic convergence.
- For the central beta distribution the algorithm is equally efective.
- For the noncentral gamma distribution, the improvement over existing algorithms (based on the secant method) will be considerable (the Bessel function I will be needed). Same for the non-central beta.
- The applicability of the method is not restricted to CDFs or to functions with negative Schwarzian derivative. Another example of application is the inversion of elliptic integrals, like

$$f(x) = \int_0^x \sqrt{1 - m^2 \sin^2 t}\, dt - p,\ 0 < m < 1$$

for which 10^{-25} relative accuracy is obtained in 2 iterations (and 10^{-40} if $m < 0.8$). A previous method by Boyd (2012) gave 10^{-10} in 3 iterations.

THANK YOU FOR YOUR ATTENTION