Computer Arithmetic

Charles J. Geyer
School of Statistics
University of Minnesota

Stat 8054 Lecture Notes

IEEE Arithmetic

What is for short called "IEEE arithmetic" is the standard for floating point numbers in nearly all currently manufactured computers (everything except IBM mainframes).

What you need to know about IEEE arithmetic is that there are two kinds of floating point numbers. In C the types are

- `float` about 6.9 decimal digits precision
- `double` about 15.6 decimal digits precision

In R only `double` is used. Go and do likewise.

IEEE Arithmetic (cont.)

IEEE arithmetic also represents values that are not ordinary floating point numbers. In R these are printed

- `NaN` meaning not a number
- `Inf` meaning $+\infty$
- `-Inf` meaning $-\infty$

IEEE Arithmetic (cont.)

These follow obvious rules of arithmetic

$$\begin{aligned}
\texttt{NaN} + x &= \texttt{NaN} \\
\texttt{NaN} * x &= \texttt{NaN} \\
\texttt{Inf} + x &= \texttt{Inf}, && x > \texttt{-Inf} \\
\texttt{Inf} + \texttt{-Inf} &= \texttt{NaN} \\
\texttt{Inf} * x &= \texttt{Inf}, && x > 0 \\
\texttt{Inf} * 0 &= \texttt{NaN} \\
x/0 &= \texttt{Inf}, && x > 0 \\
0/0 &= \texttt{NaN}
\end{aligned}$$

Overflow

In R the function `is.finite` tests that numbers are not any of `NA`, `NaN`, `Inf`, `-Inf`.

Can have `all(is.finite(x))` equal to `TRUE` but `sum(x)` or `prod(x)` equal to `Inf`. This is called overflow.

To be avoided if at all possible. Loss of all significant figures.

Example: `log(exp(710))` is `Inf` not 710.

Underflow

An IEEE arithmetic result can be zero, when the exact infinite precision result would be positive but smaller than the smallest positive number representable in IEEE arithmetic. This is called underflow.

Example: `log(exp(-746))` is `-Inf` not 746.

Underflow is not a worry if the result is later added to a large number.

Example: `log(1 + exp(-746))` is 0, which is correct.

Denormalized Numbers

Between the smallest positive number representable with full (15.6 decimal digit) precision and zero are numbers representable with less precision.

Example: `log(exp(-743))` is `-743.0538` not 743.

Catastrophic Cancellation

We say “catastrophic cancellation” occurs when subtracting two nearly equal positive numbers gives a number with much less precision.

Example

$$1.020567 - 1.020554 = 1.3 \times 10^{-5}$$

Both operands have 7 decimal digits of precision. The result has 2.

Short-Cut Formula for Variance

Never use

$$\text{var}(X) = E(X^2) - E(X)^2$$

It is an invitation to catastrophic cancellation.

Always use the two-pass algorithm

$$\begin{aligned} \bar{x}_n &= \frac{1}{n} \sum_{i=1}^{n} x_i \\ v_n &= \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \end{aligned}$$

There is also sophisticated one-pass algorithm (see links).

A Problem Requiring Care

The log likelihood for the binomial distribution is

$$l(p) = x \log(p) + n \log(1 - p)$$

In terms of the natural parameter

$$\theta = \operatorname{logit}(p) = \log\left(\frac{p}{1 - p}\right)$$

$$p = \frac{e^{\theta}}{1 + e^{\theta}}$$

the log likelihood is

$$l(\theta) = x\theta - n \log(1 + e^{\theta})$$

A Problem Requiring Care (cont.)

From the properties of likelihood

$$\begin{aligned} E_\theta\{\nabla l(\theta)\} &= 0 \\ \mathrm{var}_\theta\{\nabla l(\theta)\} &= -E_\theta\{\nabla^2 l(\theta)\} \end{aligned}$$

get

$$\begin{aligned} l'(\theta) &= x - E_\theta(x) \\ &= x - np(\theta) \\ l''(\theta) &= -\mathrm{var}_\theta(x) \\ &= -np(\theta)q(\theta) \end{aligned}$$

where $q(\theta) = 1 - p(\theta)$.

Log Likelihood Function Itself

For case $\theta \leq 0$ formula

$$l(\theta) = x\theta - n \log(1 + e^{\theta})$$

is well behaved (no overflow), otherwise

$$\begin{aligned} l(\theta) &= x\theta - n \log\bigl(e^{\theta}(e^{-\theta} + 1)\bigr) \\ &= (x - n)\theta - n \log(1 + e^{-\theta}) \end{aligned}$$

is well behaved.

In both cases we should use the function

$$\text{log1p}(x) = \log(1 + x)$$

(defined in both C99 and R) for more accurate calculation when x is near zero.

Probabilities

$$
\begin{aligned}
p(\theta) &= \frac{1}{1 + \exp(-\theta)} \\
q(\theta) &= \frac{1}{1 + \exp(\theta)}
\end{aligned}
$$

Suffer from neither overflow nor catastrophic cancellation (if a denominator overflows, the result is zero, which is correct).

Note well:

$$q(\theta) = 1 - p(\theta)$$

can suffer from catastrophic cancellation. Don't ever do that!

Derivatives

$$l'(\theta) = x - np(\theta) \tag{a}$$
$$l''(\theta) = -np(\theta)q(\theta) \tag{b}$$

are fine when $p(\theta)$ and $q(\theta)$ are calculated properly.

(a) does suffer from catastrophic cancellation when the result is nearly zero, but there appears to be no remedy.

Since p is a smooth function,

$$p'(\theta) = p(\theta)q(\theta)$$

this limits the accuracy of the solution to more or less the accuracy of computer arithmetic.

Another Problem Requiring Care

A Monte Carlo approximation of the log likelihood for an exponential family with unknown normalizing constant is

$$l_n(\theta) = \langle x, \theta \rangle - \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\langle x_i, \theta - \psi \rangle} \right)$$

where x is the observed data, θ a free variable ranging over the parameter space, $\langle x, \theta \rangle$ denotes the inner product of vectors (write $x^T \theta$ if you prefer), and x_1, x_2, $\ldots$ are simulations from the distribution in the family with parameter vector ψ.

Another Problem Requiring Care (cont.)

$$l_n(\theta) = \langle x, \theta \rangle - \log\left(\frac{1}{n}\sum_{i=1}^{n} e^{\langle x_i, \theta - \psi \rangle}\right)$$

The exponentials can overflow to `+Inf` when the whole expression is reasonably sized.

Two-Pass Algorithm for the Function Itself

Let

$$a = \max_{1 \le i \le n} \langle x_i, \theta - \psi \rangle$$

Then

$$\begin{aligned} l_n(\theta) &= \langle x, \theta \rangle - \log\left(\frac{e^a}{n} \sum_{i=1}^{n} e^{\langle x_i, \theta - \psi \rangle - a} \right) \\ &= \langle x, \theta \rangle - a - \log\left(\frac{1}{n} \sum_{i=1}^{n} e^{\langle x_i, \theta - \psi \rangle - a} \right) \end{aligned}$$

Overflow problem fixed.

Two-Pass Algorithm for the Function Itself (cont.)

Let

$$b = a - \langle x, \theta - \psi \rangle$$

Then

$$\begin{aligned} l_n(\theta) &= \langle x, \theta \rangle - a - \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\langle x_i, \theta - \psi \rangle - a} \right) \\ &= \langle x, \psi \rangle - b - \log \left(\frac{1}{n} \sum_{i=1}^{n} e^{\langle x_i - x, \theta - \psi \rangle - b} \right) \end{aligned}$$

Cancellation problems improved. $x_i - x$ tends to be small when ψ is near MLE and approximation is only good when θ is near ψ.

First Derivative

$$\nabla l_n(\theta) = -\frac{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}(x_i - x)e^{\langle x_i - x, \theta - \psi\rangle - b}}{\dfrac{1}{n}\displaystyle\sum_{i=1}^{n} e^{\langle x_i - x, \theta - \psi\rangle - b}}$$

Second Derivative

$$\nabla^2 l_n(\theta) = -\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)(x_i - x)^T e^{\langle x_i - x, \theta - \psi\rangle - b}}{\frac{1}{n}\sum_{i=1}^{n} e^{\langle x_i - x, \theta - \psi\rangle - b}}$$

$$+ \left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)e^{\langle x_i - x, \theta - \psi\rangle - b}}{\frac{1}{n}\sum_{i=1}^{n} e^{\langle x_i - x, \theta - \psi\rangle - b}}\right)\left(\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - x)e^{\langle x_i - x, \theta - \psi\rangle - b}}{\frac{1}{n}\sum_{i=1}^{n} e^{\langle x_i - x, \theta - \psi\rangle - b}}\right)^T$$

Simplification

$$b = \max_{1 \le i \le n} \langle x_i - x, \theta - \psi \rangle$$

$$w_i = \frac{e^{\langle x_i - x, \theta - \psi \rangle - b}}{\sum_{j=1}^{n} e^{\langle x_j - x, \theta - \psi \rangle - b}}$$

$$\nabla l_n(\theta) = -\sum_{i=1}^{n} (x_i - x) w_i$$

$$\nabla^2 l_n(\theta) = -\sum_{i=1}^{n} (x_i - x)(x_i - x)^T w_i + [\nabla l_n(\theta)][\nabla l_n(\theta)]^T$$

Avoiding Catastrophic Cancellation

$$s_\theta = \nabla l_n(\theta)$$

$$\sum_{i=1}^{n} (x_i - x + s_\theta) w_i = 0$$

$$\begin{aligned} -\sum_{i=1}^{n} (x_i - x + s_\theta)(x_i - x + s_\theta)^T w_i &= -\sum_{i=1}^{n} (x_i - x)(x_i - x)^T w_i \\ &\qquad - 2 s_\theta \sum_{i=1}^{n} (x_i - x)^T w_i - s_\theta s_\theta^T \\ &= -\sum_{i=1}^{n} (x_i - x)(x_i - x)^T w_i + s_\theta s_\theta^T \\ &= \nabla^2 l_n(\theta) \end{aligned}$$

Recap

$$u_i = \langle x_i - x, \theta - \psi \rangle$$

$$b = \max_{1 \le i \le n} u_i$$

$$v_i = \exp(u_i - b)$$

$$l_n(\theta) = \langle x, \psi \rangle - b - \log \left(\frac{1}{n} \sum_{i=1}^{n} v_i \right)$$

$$w_i = v_i \Big/ \sum_{i=1}^{n} v_i$$

$$s_\theta = \nabla l_n(\theta) = - \sum_{i=1}^{n} (x_i - x) w_i$$

$$\nabla^2 l_n(\theta) = - \sum_{i=1}^{n} (x_i - x + s_\theta)(x_i - x + s_\theta)^T w_i$$