Fast Magnitude Calculation

By Clay S. Turner
5/27/09 V1.2

Often one needs to find the magnitude of a two dimensional vector quickly without using a square root function. A common approximation takes the following form:

$$|r| \approx \alpha \cdot \max\{|x|,|y|\} + \beta \cdot \min\{|x|,|y|\} \qquad [1]$$

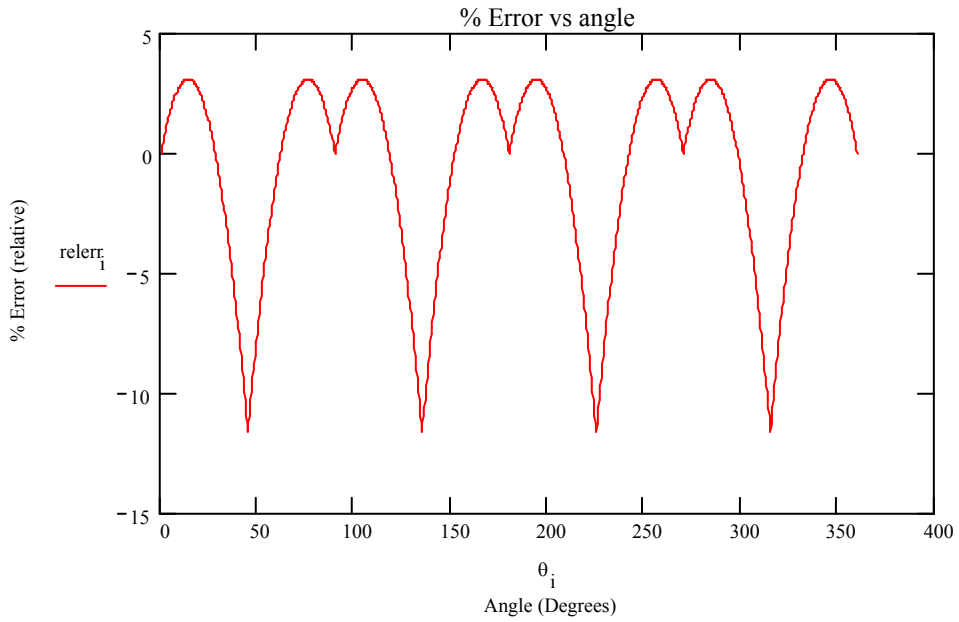
Where x and y are the two components of the vector and α and β are constants. The trick is to determine the two constants so as to make the approximation useful.

This has been talked about on the web a fair bit as this is a useful trick.

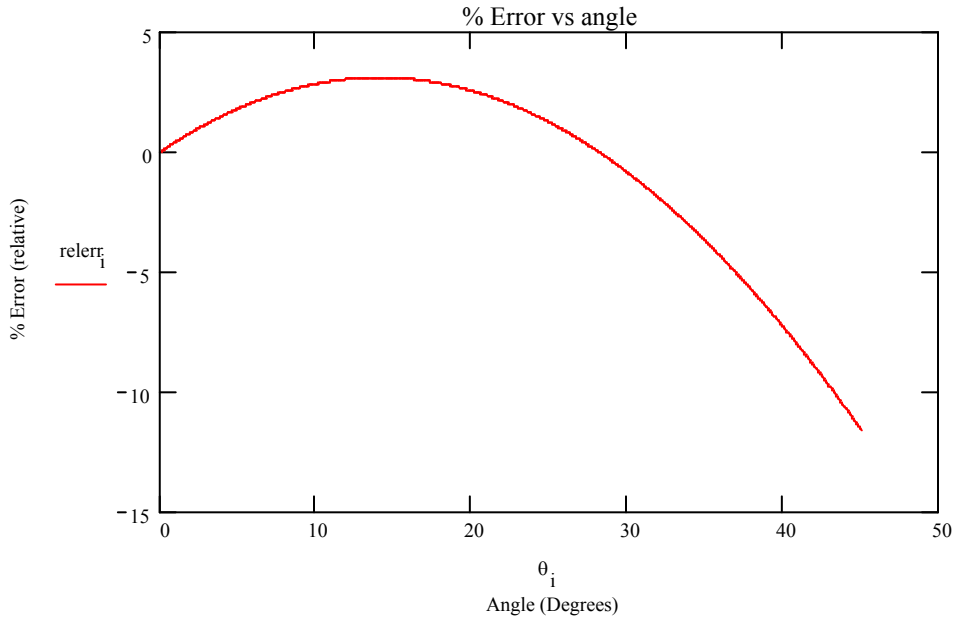
See http://www.dspguru.com/comp.dsp/tricks/alg/mag_est.htm for Grant Griffin's description. As I provided the coefficients for the "optimized" cases in his write up, I thought I'd put together some of the details of the math behind how these magic values are found for this problem.

Some may question the wisdom of going to such lengths (like finding the precise values for the constants) for something that is inherently low precision. To be sure when this trick was 1st being used, strong incentives existed to minimize hardware and thus using low precision for the coefficients was acceptable. In fact they were often chosen for having few bits in binary representation. But nowadays, high precision multipliers abound, so using a coefficient with a lot of decimal places causes little to no extra computational effort. Plus sometimes it is good to know what the limits of the method are!

A crude but computationally simple set is $\alpha = 1$ and $\beta = 0.25$. The following is a graph of the relative error for this case:



As is evident from this graph, the error function repeats every 90 degrees with a reflection about the center point of each 90 degree region. Thus the error function needs to be only examined over a range of 45 degrees as this contains all of the details for the whole 360 degree range. The same function over the reduced domain follows:



Here the relative error is observed to range from +3.078% down to -11.612%.

There are several methods of optimization that we will look at. First we will use the concept of least squared error. Essentially this means the total of the square of the error is minimized.

How do we define the error? Relative error is commonly defined as being the (approximate formulation minus the exact formulation) divided by the exact formulation. The relative error function for the fast magnitude approximation may be written as follows:

$$err = \frac{\alpha \cdot r\cos(\theta) + \beta \cdot r\sin(\theta) - r}{r} \qquad [2]$$

Here we have taken a vector (magnitude r and angle theta) and then put it (as resolved into its two Cartesian components) into [1]. Since the x component is greater than or equal to the y component over the range of 0 to 45 degrees, [1] reduces to a simpler form without the "max" and "min" functions. Then [1]'s result is placed into the standard relative error formula.

The squared error is simply:

$$err^2 = \left[\alpha \cdot \cos(\theta) + \beta \cdot \sin(\theta) - 1\right]^2 \qquad [3]$$

The total squared error is the integral of [3] over the range of 0 to 45 degrees.

$$\int_0^{\frac{\pi}{4}} \left(\alpha \cdot \cos(\theta) + \beta \cdot \sin(\theta) - 1\right)^2 d\theta \qquad [4]$$

The result is:

$$\psi = \left(\frac{\pi}{8} + \frac{1}{4}\right)\alpha^2 - \sqrt{2}\alpha + \left(\frac{\pi}{8} - \frac{1}{4}\right)\beta^2 + \left(\sqrt{2} - 2\right)\beta + \frac{\alpha\beta}{2} + \frac{\pi}{4} \qquad [5]$$

And then to minimize this, we then find out where the gradient (with respect to our two unknown constants) of the total squared error is zero[1]. Thus we need to find out when:

$$\vec{\nabla} \int_0^{\pi/4} \left[\alpha \cdot \cos(\theta) + \beta \cdot \sin(\theta) - 1\right]^2 d\theta = 0 \qquad [6]$$

This vector equation may be written as two scalar equations:

[1] Specifically by equating this to zero, we are finding either a maximum or a minimum or a saddle point. By examining the 2nd order derivatives or a detailed enough plot, we can ascertain the type of point.

$$\frac{\partial}{\partial\alpha}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\left(\frac{1}{2}+\frac{\pi}{4}\right)\alpha+\frac{1}{2}\beta-\sqrt{2}=0 \qquad [7]$$

And

$$\frac{\partial}{\partial\beta}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{1}{2}\alpha+\left(\frac{\pi}{4}-\frac{1}{2}\right)\beta+\sqrt{2}-2=0 \qquad [8]$$

Next one uses the results of [7] and [8] and finds the simultaneous solution for alpha and beta. It is:

$$\alpha=\frac{4\left(\pi\sqrt{2}-4\right)}{\pi^2-8}\approx 0.947543636290784 \qquad [9]$$

$$\beta=\frac{4\left(4+2\sqrt{\pi}-(4+\pi)\sqrt{2}\right)}{\pi^2-8}\approx 0.392485425091961 \qquad [10]$$

To verify the quality (type of) solution, we just find the 2nd order derivatives of [7] and [8]. They are:

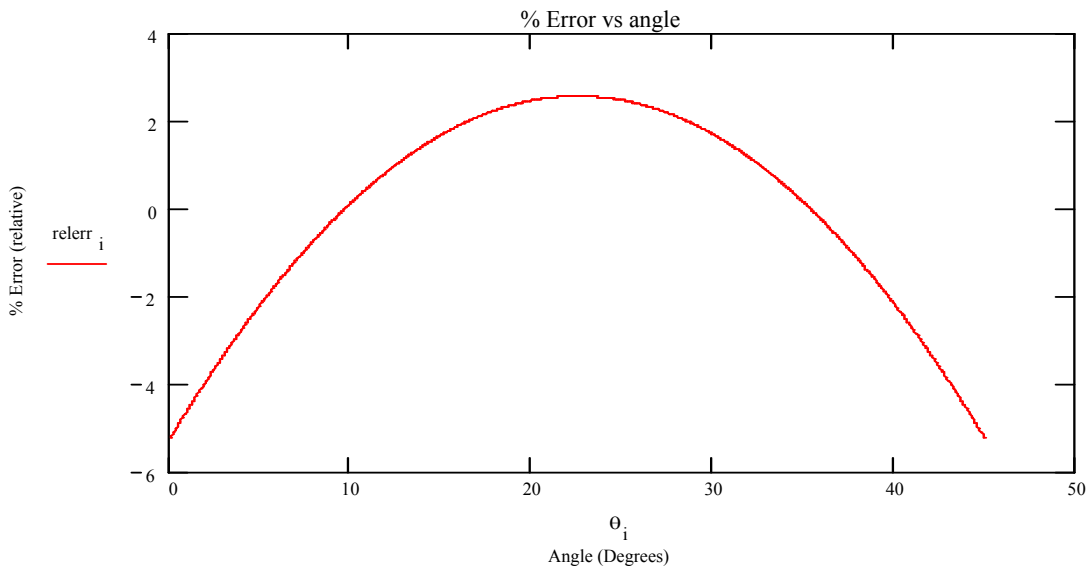
$$\frac{\partial^2}{\partial\alpha^2}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{1}{2}+\frac{\pi}{4} \qquad [11]$$

And

$$\frac{\partial^2}{\partial\beta^2}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{\pi}{4}-\frac{1}{2} \qquad [12]$$

Since [11] and [12] are always >0, then a solution to [6], if it exists (and it does in this case), is a global minimum. Thus, [9] and [10] are the coefficients that yield the least squared error to [6].

If we use the values from [9] and [10] in [1], we get the following relative error function:



These coefficients certainly reduce the error when compared to when the crude set is used. Our error now ranges from +2.561% down to -5.246%. Even though this is the best that using least squared error can give us, some will note that this has a nonzero average error. This may cause an unnecessary bias in some computations. The average error is given by the following:

$$\mu = \frac{4}{\pi}\int_0^{\pi/4}(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1)d\theta = \frac{2\sqrt{2}\cdot\alpha+(4-2\sqrt{2})\cdot\beta-\pi}{\pi} \qquad [13]$$

Our particular coefficients [9], and [10] inserted into [13] from the least squared error optimization results in the average error of:

$$\mu = \frac{-\left(128\left(\sqrt{2}-1\right)+32\pi\sqrt{2}-72\pi+\pi^3\right)}{\pi\left(\pi^2-8\right)} \approx -0.000544072081 = -0.0544072081\%$$

While small it is not zero. It works out to be about one one-hundredth of the maximum error. This smallness suggests that constraining the average error to be zero may result in a solution not very different from the unconstrained version.

So let's now find the coefficients resulting from minimizing the least squared error subject to the constraint of zero average. For this we use the method of Lagrange multipliers.

Lagrange's method states that points for where the gradient of the function to be optimized is parallel to the gradient of the constraining function, that those points are stationary. So we need to find that point and be sure it is a minimum! For our case with least squared error and a zero average, the vector equation is:

$$\vec{\nabla}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\vec{\nabla}\frac{4\lambda}{\pi}\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta \qquad [14]$$

Unlike before, this vector equation is three scalar equations, so we have:

$$\frac{\partial}{\partial\alpha}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{4}{\pi}\frac{\partial}{\partial\alpha}\lambda\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta \qquad [15]$$

$$\frac{\partial}{\partial\beta}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{4}{\pi}\frac{\partial}{\partial\beta}\lambda\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta \qquad [16]$$

$$\frac{\partial}{\partial\lambda}\int_0^{\pi/4}\left[\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right]^2 d\theta=\frac{4}{\pi}\frac{\partial}{\partial\lambda}\lambda\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta \qquad [17]$$

We already know the left hand integrals (the first two anyway and the third is trivial). The right hand sides of [15], [16], and [17] are:

$$\frac{4\lambda}{\pi}\frac{\partial}{\partial\alpha}\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta=\frac{\lambda 2\sqrt{2}}{\pi} \qquad [18]$$

$$\frac{4\lambda}{\pi}\frac{\partial}{\partial\beta}\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta=\frac{\lambda 2\left(2-\sqrt{2}\right)}{\pi} \qquad [19]$$

$$\frac{4}{\pi}\frac{\partial}{\partial\lambda}\lambda\int_0^{\pi/4}\left(\alpha\cdot\cos(\theta)+\beta\cdot\sin(\theta)-1\right)d\theta=\frac{2\sqrt{2}(\alpha-\beta)-\pi+4\beta}{\pi} \qquad [20]$$

Thus [15], [16], and [17] reduce to:

$$\left(\frac{1}{2}+\frac{\pi}{4}\right)\alpha+\frac{1}{2}\beta-\sqrt{2}=\frac{\lambda 2\sqrt{2}}{\pi} \qquad [21]$$

$$\frac{1}{2}\alpha+\left(\frac{\pi}{4}-\frac{1}{2}\right)\beta+\sqrt{2}-2=\frac{\lambda 2\left(2-\sqrt{2}\right)}{\pi} \qquad [22]$$

$$0=\frac{2\sqrt{2}(\alpha-\beta)-\pi+4\beta}{\pi} \qquad [23]$$

The simultaneous solution to [21], [22], and [23] results in:

$$\alpha = \frac{\pi}{8}\left(1+\sqrt{2}\right) \approx 0.948059448968522 \qquad [24]$$

$$\beta = \frac{\pi}{8} \approx 0.392699081698724 \qquad [25]$$

And as we expected the values [24] and [25] are very close to the values [9] and [10].

Next we will look at a min-max solution. This name stems from the concept of minimizing the maximum error of the approximation over the interval of interest. This type of optimization results in an equiripple error (difference between the approximated and desired functions). For a 1st order approximation, our error function will have 3 extrema, all with identical magnitudes. The extrema also will have alternating signs as the error function is oscillatory. Since we are approximating a smooth function over a compact interval, then our endpoints are extrema. Thus we need to just find the 3rd extremal point.

Let's define the deviation between the approximation and the ideal function at our extremal points to have a value of rho. The ideal magnitude function for a normalized vector is unity. And we will use the two endpoints of our domain and a third intermediate point within the domain, thus giving the following matrix formulation for our problem is:

$$\begin{bmatrix} 1 & 0 & 1 \\ \cos(\theta) & \sin(\theta) & -1 \\ \cos(\pi/4) & \sin(\pi/4) & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \rho \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \qquad [26]$$

Remez's[2] theory of Chebyshev approximation says that when the magnitude of rho is maximized (as the abscissal values are moved about), then we have the equi-ripple solution. So solving [26] we find for rho:

[2] In this case I'm exploiting a very simplified case where only one point is movable – the other two are locked at the endpoints of the domain. In general for higher order cases, an iterative procedure (Remez Exchange Algorithm) is used to maximize the magnitude of rho and locate the corresponding optimizing abscissal values from which the approximating polynomial is formed..

$$\rho(\theta) = \frac{2\sqrt{2-\sqrt{2}}\cos\left(\theta - \pi/8\right) - \sqrt{2}}{2\sqrt{2-\sqrt{2}}\cos\left(\theta - \pi/8\right) + \sqrt{2}} \qquad [27]$$

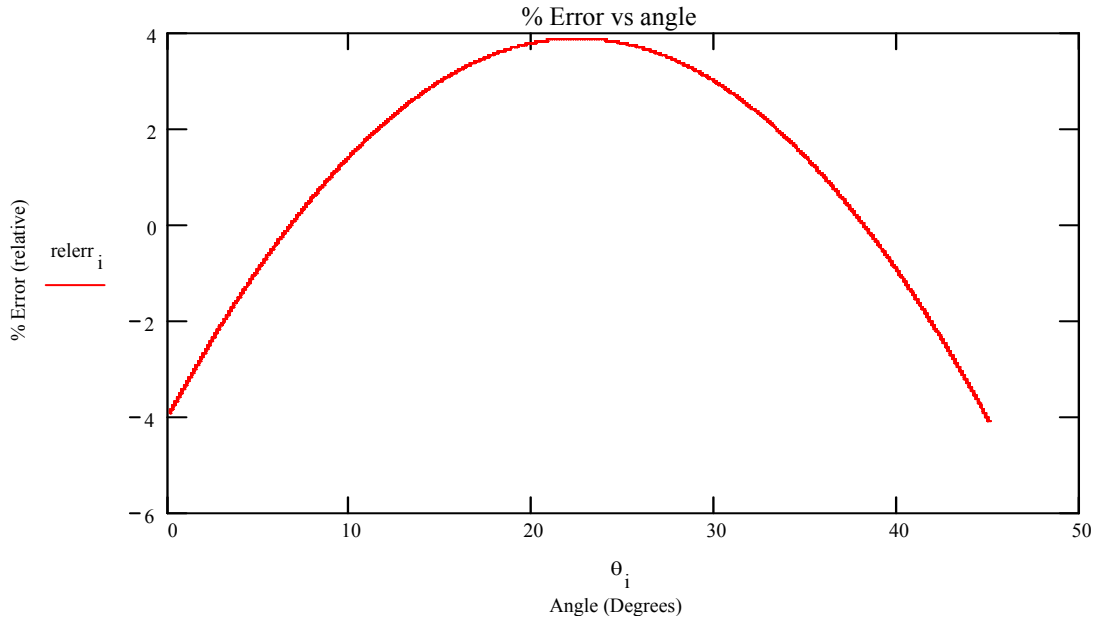
Taking the derivative of [27] and finding when it is zero, yields $|\rho|$ is maximized when $\theta = \pi/8$ So putting this value into [26] and solving, we find:

$$\alpha = \frac{2\sqrt{2}}{2\sqrt{2-\sqrt{2}} + \sqrt{2}} \approx 0.96043387010342 \qquad [28]$$

$$\beta = \frac{4 - 2\sqrt{2}}{2\sqrt{2-\sqrt{2}} + \sqrt{2}} \approx 0.397824734759316 \qquad [29]$$

$$\rho = \frac{2\sqrt{2-\sqrt{2}} - \sqrt{2}}{2\sqrt{2-\sqrt{2}} + \sqrt{2}} \approx 0.03956612989658 \qquad [30]$$

The following plot shows the equi-ripple error function resulting from using [28] and [29] in [1]. The magnitude of the maximum error is given by [30] which we can see gives a maximum relative error of just under 4%.



In summary we have the following three sets of coefficients depending on the type of optimization involved:

Optimization	Alpha		Beta	
Least Squared Error	$\frac{4(\pi\sqrt{2}-4)}{\pi^2-8}$	0.947543	$\frac{4\left(4+2\sqrt{\pi}-(4+\pi)\sqrt{2}\right)}{\pi^2-8}$	0.392485
Least Squared Error with Zero Mean	$\frac{\pi(4-\pi\sqrt{2})}{8((4+\pi)\sqrt{2}-4-2\pi}$	0.948059	$\frac{\pi}{8}$	0.392699
Equiripple Error	$\frac{2\sqrt{2}}{2\sqrt{2-\sqrt{2}}+\sqrt{2}}$	0.960434	$\frac{4-2\sqrt{2}}{2\sqrt{2-\sqrt{2}}+\sqrt{2}}$	0.397825