TABULATION OF SATURATION MODELS

Dragan Vidović*, Milan Dotlić*, Milan Dimkić* and Milenko Pušić†

*Institute for the Development of Water Resources Jaroslav Černi
Belgrade, Serbia
e-mail: dragan.vidovic@jcerni.co.rs

†Faculty of Mining and Geology
Belgrade University, Serbia

Key words: Saturation models, Van Genuchten, tabulation

Summary. Tabulation of saturation models such as Van Genuchten-Mualem formulae requires a compromise between the memory consumption and the computational cost. We present a tabulation method based on the binary representation of a floating point number which evaluates quickly and has modest memory requirements.

1 INTRODUCTION

Evaluation of saturation and relative conductivity using analytical models such as Van Genuchten and Mualem formulae is computationally expensive because it requires evaluation of power functions with fractional exponent. For this reason, these models are tabulated and evaluated approximately using splines.

Solution of Richards equation can be quite sensitive to the relative conductivity evaluation accuracy, requiring short spline sub-intervals, higher-order splines, or both. Pressure head variation can be large or not known in advance, for example if the unsaturated zone is large. This variation must be estimated and divided in short sub-intervals for spline interpolation. Such a table may require much memory.

An alternative is to use non-uniform sub-intervals. Relative conductivity changes rapidly with the pressure head in a narrow region of higher saturation, and more slowly in drier conditions. Thus using short sub-intervals near the saturation point, and longer sub-intervals away from it, saves the memory and provides the accuracy. However, mapping the pressure head into non-uniform sub-intervals may require expensive evaluation of analytical functions, potentially annihilating the speedup achieved by the tabulation.

In order to conserver the memory and speed up the computation, we propose such non-uniform division where mapping to sub-intervals is not much more expensive than in the uniform case.
2 BACKGROUND

Darcy’s law extended to the multiphase case can be written as

\[
\mathbf{u} = -k_r \frac{k}{\mu} (\nabla p - \rho g),
\]

where \( \mathbf{u} \) is the Darcy velocity, \( k_r \) is the relative conductivity, \( k \) is the hydraulic permeability tensor of the saturated medium, \( \mu \) is the dynamic viscosity, \( p \) is the water pressure, \( \rho \) is the water density, and \( g \) is the gravity vector pointing downwards.

Substituting Darcy’s law into the continuity equation

\[
\frac{\partial \theta}{\partial t} = -\nabla \cdot \mathbf{u}
\]

results in the so-called mixed formulation of Richards equation

\[
\frac{\partial \theta}{\partial t} = \nabla \cdot \left[ k_r \frac{k}{\mu} (\nabla p - \rho g) \right],
\]

where \( \theta = \varepsilon s \) is the water content, \( \varepsilon \) is the porosity, \( s \) is the saturation, and \( t \) is the time. Atmospheric pressure \( p_0 \) is assumed to be constant, which eliminates the equation for the air phase pressure.

The saturation and the relative conductivity are usually computed from the pressure head \( \psi = (p - p_0)/(\rho g) \) using the Van Genuchten-Mualem model [4, 3]:

\[
s_e = \begin{cases} 
\frac{1}{(1 + |\alpha \psi|^n)^m} & \text{if } \psi < \psi_a \\
1 & \text{if } \psi \geq \psi_a 
\end{cases}
, \quad s_e = \frac{s - s_r}{s_s - s_r},
\]

\[
k_r = s_e^{\frac{1}{2}} \left[ 1 - \left( 1 - s_e^{\frac{1}{m}} \right)^{m^2} \right],
\]

where \( s_e \) is the effective saturation, \( s_r \) is the residual saturation, \( s_s \) is the maximal saturation (usually \( s_s = 1 \)), \( \alpha \) is a curve-fitting parameter, \( n \geq 1 \) is the pore size distribution index, \( m = 1 - 1/n \), and \( \psi_a \) is the air-entry pressure head. Other models, like Brooks-Corey [1], may be used instead.

3 TABULATION

Functions \( s_e(\psi) \) and \( k_r(\psi) \) are shown in Fig 1. Their analytical expressions (4), (5) involve expensive power functions. To reduce the computation cost, they are tabulated and evaluated using linear or higher-order splines [2, 5]. Both functions change rapidly near the saturation point, and very slowly in drier conditions. Thus, it would be advantageous to interpolate them with finer resolution when \( \psi \to 0 \).

Assuming that function \( f(x) \) defined for \( x > 0 \) changes rapidly when \( x < x_0 = 2^{M_0} \) and slows down further away from zero, we use uniform sub-intervals of length \( l_{M_0} = \ldots \)
\[ \frac{2^{M_0}}{(2N)} \text{ for tabulation in } (0, 2^{M_0}), \text{ and uniform sub-intervals of length } l_M \text{ for tabulation in } [2^{M-1}, 2^M], \ M > M_0 \text{ (see Fig. 2). The key for the efficiency is that one does not have to evaluate a logarithmic function in order to find } M \text{ such that } x \in [2^{M-1}, 2^M). \text{ This information is available in the binary representation of a positive floating point number} \]

\[ x = a \cdot 2^M, \]

where \( a \) is the significand such that \( 0.5 \leq a < 1 \).

Exponent \( M \) could be extracted by aliasing \( x \) to an integer and then performing a few cheap bitwise operations. However, this would break the strict aliasing rules, and such code is unlikely to work if compiler optimization is used.

Fortunately, there are legal ways to break up a floating point number: in standard C this is achieved with function \texttt{frexp}, while in FORTRAN 95 and later intrinsic functions \texttt{EXponent} and \texttt{FRACTION} perform the same task.

Function \( f(x) \) is tabulated at points \( x^{M_0}_j = j l_{M_0}, \ 0 \leq j < 2N, \) and at points \( x^M_j = 2^{M-1} + j l_m, \ 0 \leq j < N, \ M_0 < M < M_{\text{max}}. \) We denote \( f^M_j = f(x^M_j). \) Derivatives \( d^M_j = \frac{f^M_{j+1} - f^M_j}{l_M} \) are also computed. The following algorithm is used to evaluate the linear spline:
\[
\begin{align*}
\text{if } x \leq x_0 \text{ then} \\
j &= \lfloor x/l_{M_0} \rfloor \\
\text{return } f^j_{M_0} + (x - x^j_{M_0})d^j_{M_0} \\
\text{else} \\
\text{Extract the exponent } M \text{ and the significand } a \text{ such that (6) holds} \\
j &= \lfloor (a - 0.5)/(2N) \rfloor \\
\text{return } f^j_{M} + (x - x^j_{M})d^j_{M} \\
\end{align*}
\]

4 Example

We compute \( k_r(\psi) \) with parameters \( \alpha = 4.1, n = 1.964, s_s = 1.0, s_r = 0.0025 \) between \(-2^{13}\) and 0 using the described scheme. We take \( f(-\psi) = k_r(\psi), x_0 = 0.25 \) and \( N = 2^8 \).

![Figure 3: Relative error of \( k_r \) plotted against \( -\psi \).](image)

Figure 3 shows the approximation relative error in sub-interval central points. The relative error order of magnitude does not change as \( \psi \to \infty \). For large \(-\psi\) the error behavior is less regular due to round-off errors, because in this zone \( k_r \) is almost zero. The choice of \( x_0 \) is such that the relative errors in the linear and in the exponential part are of similar order of magnitude. Except for the zone near the inflection point of \( k_r \), there is no region in which the function is evaluated much more accurately than elsewhere, which indicates that this spline uses the memory for sampling points in a way which is close to optimal.
5 CONCLUSIONS

A method for choosing interpolation intervals for a linear spline appropriate for evaluation of the relative conductivity and the saturation from the pressure head has been presented. This method puts together the low evaluation computational effort of uniformly distributed intervals with the memory efficiency of non-uniformly distributed intervals.

Higher-order spline interpolation can also be formulated for this interval distribution.

REFERENCES


