How to Floating Point Quantize

To convert from a number into a scale-mantissa floating point code with Rs scale bits and Rm mantissa bits:

I. Quantize the number as an R bit code where $R=2^{Rs}-1+Rm$.

II. Count the number of leading zeros in |code|. If the number of leading zeros is less than $2^{Rs}-1$ then set the scale equal to the number of leading zeros; otherwise set the scale equal to $2^{Rs}-1$.

III. If scale equals $2^{Rs}-1$, then set the first mantissa bit equal to s and set the remaining Rm-1 bits equal to the bits following the $2^{Rs}-1$ leading zeros in |code|; otherwise set the first mantissa bit equal to s and set the remaining Rm-1 bits equal to the bits following the leading zeros omitting the leading one.

How to Floating Point De-Quantize

To convert from scale-mantissa floating point code with Rs scale bits and Rm mantissa bits into a number:

I. Create an R bit code where $R=2^{Rs}-1+Rm$ from the mantissa and scale factor where s is the first mantissa bit and |code|
- **A. has scale leading zeros**
- **B. followed by the remaining Rm-1 mantissa bits if scale is $2^{Rs}-1$, otherwise followed by a one and then the remaining mantissa bits**
- **C. followed by a one and as many trailing zeros as will fit if scale is less than $2^{Rs}-1$.**

II. Dequantize the R bit code into the number.