5.19 Let $E(e^{j\omega T})$ be the error function caused by rounding the coefficients. We have:

$$E(e^{j\omega T}) = e^{-j\omega(K+1)T}\,[\delta h_0 + 2\sum_{n=1}^{K} \delta h_n\, cos(\omega n T)]$$

but $|\delta h_n| \le Q_c/2$. The deviation can be estimated in may ways, for example, the maximal deviation or the variance of the deviation. Here we use an estimate of the maximal deviation:

$$|E(e^{j\omega T})| = |e^{-j\omega(K+1)T}\,[\delta h_0 + 2\sum_{n=1}^{K} \delta h_n\, cos(\omega n T)]| \le$$

$$\le |\delta h_0 + 2\sum_{n=1}^{K} \delta h_n\, cos(\omega n T)| \le |\delta h_0| + 2\sum_{n=1}^{K} |\delta h_n\, cos(\omega n T)| \le$$

$$|E(e^{j\omega T})| \le |\delta h_0| + 2\sum_{n=1}^{K} |\delta h_n|$$

Hence, $|E(e^{j\omega T})| \le \frac{Q}{2}(1 + 2K) = \frac{NQ}{2}$ (Bound #1)

Now, assume that the coefficients are randomly rounded and the errors are considered as independent random variables that are uniformly distributed in the interval $[-Q/2, Q/2]$. The variance is $Q^2/12$. Let e_0 be the effective value of $E(e^{j\omega T})$ in the passband:

$$e_0{}^2 = \frac{1}{f_c}\int_0^{f_c} |E(e^{j\omega T})|^2 d\omega T = \sum_{n=0}^{N-1} |\delta h_0|^2$$

The variance of $e_0{}^2$ is:

$$\sigma^2 = E\{e_0{}^2\} = \frac{NQ^2}{12} \qquad \text{(Bound \#2)}$$

The required coefficient word length is estimated as follows. Let δ_m be the acceptable deviation in the passband of the stopband. Now, we must have:

$$|E(e^{j\omega T})| < \delta_m - \delta_0$$

where δ_0 is the deviation before quantization of the coefficients. Let the level of acceptance of a coefficient set that does not fit the requirements be 5%. Hence the probability the coefficient set does not meet the specification is (assuming a normal distribution):

$$P(|E(e^{j\omega T})| \ge 2\sigma_x) = P(\frac{|E(e^{j\omega T})|}{\sigma_x} \ge 2) \approx 0.05$$

$$2[1-\Phi\left(\frac{|E(e^{j\omega T})|}{\sigma_x}\right)] = 0.05 \quad \Rightarrow \quad \frac{|E(e^{j\omega T})|}{\sigma_x} \approx 2$$

$$|E(e^{j\omega T})| \approx 2\sigma_x \quad \Rightarrow \quad 2\sigma_x \approx \delta_m - \delta_0$$

$$\sqrt{\frac{N\,Q^2}{12}} \approx \frac{\delta_m - \delta_0}{2} \text{ and } Q \approx (\delta_m - \delta_0)\sqrt{\frac{3}{N}}$$

The number of bits that is required to represent the coefficient depends on the largest coefficient. Hence.

$Q = 2^{(1-W_c)}\,[max\{\,[h_n|\ \} = 2^{(1-W_c)}h_0 \approx 2^{(1-W_c)}\frac{f_s + f_c}{f_{sample}}$ for a lowpass filter. Hence,

$$1 - W_c \approx log_2\left\{\frac{f_{sample}}{f_s + f_c}\,Q\right\} \approx log_2\left\{\frac{f_{sample}}{f_s + f_c}\,(\delta_m - \delta_0)\sqrt{\frac{3}{N}}\right\}$$

$$W_c \approx 1 - log_2\left\{\frac{f_{sample}}{f_s + f_c}\,(\delta_m - \delta_0)\sqrt{\frac{3}{N}}\right\}$$

For most lowpass filter we have: $\frac{N}{3} \leq \frac{f_s - f_c}{f_{sample}}$.

$$W_c \geq 1 - log_2\left\{\frac{f_{sample}}{(f_s + f_c)}\sqrt{\frac{f_{sample}}{f_s - f_c}}\,(\delta_m - \delta_0)\right\}$$

In practice, we may select $\delta_m = 2\,Min\{\,\delta_1, \delta_2\,\} = 2\delta_0$

$$W_c \geq 1 - log_2\left\{\frac{f_{sample}}{(f_s + f_c)}\sqrt{\frac{f_{sample}}{f_s - f_c}}\,\frac{2}{\delta_m}\right\}$$

$$W_c \geq 1 - log_2\left\{\frac{f_{sample}}{(f_s + f_c)}\sqrt{\frac{f_{sample}}{f_s - f_c}}\right\} + log_2\left\{\,Min\{\,\delta_1, \delta_2\,\}\right\}$$

See also: Niedringshaus W.P., Steglitz K., and Kodek D.: An Easily Computed Performance Bound for Finite Wordlength Direct-Form FIR Digital Filters, IEEE Trans. on Circuits and Systems, Vol. CAS-29, No. 3, pp. 191-193, March 1982.