Subtracting the two overlapping window sums proves the identity. It retains the same input-output map in exact arithmetic.

Comparative cost and preferred method

Direct evaluation needs $M-1$ additions per output and one scale. A rolling implementation needs one subtraction, one addition and one scale after $O(M)$ initialization, with an M-sample ring buffer. For a long stream this is $\Theta(N)$ versus $\Theta(NM)$ arithmetic. Rounding error in a recursively updated sum can accumulate, so periodic exact-window refresh or compensated accumulation is a separate numerical choice.