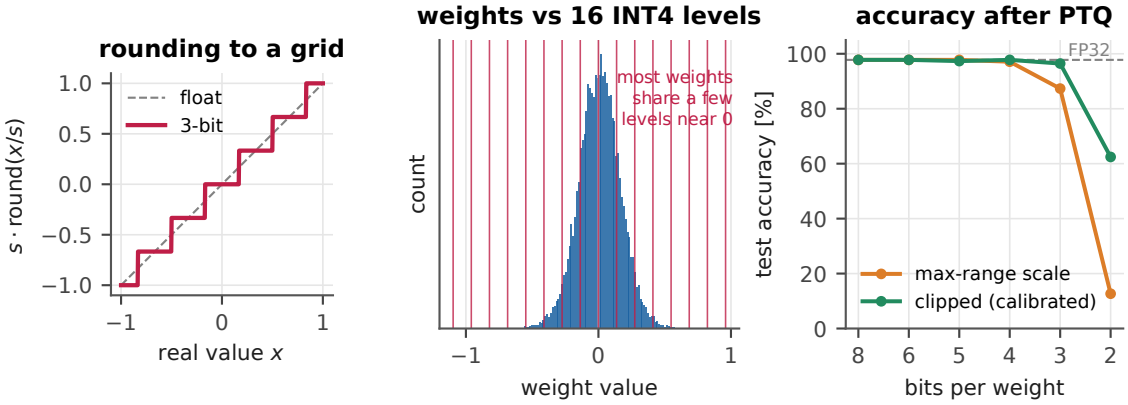

rounding to a grid
float
3-bit
1.0
0.5
0.0
−0.5
−1.0
s · round(x/s)
−1
0
1
real value x
weights vs 16 INT4 levels
most weights
share a few
levels near 0
count
−1
0
1
weight value
accuracy after PTQ
FP32
100
80
60
40
20
0
test accuracy [%]
max-range scale
clipped (calibrated)
8
6
5
4
3
2
bits per weight