Systems for Machine Learning (CMSC828G)

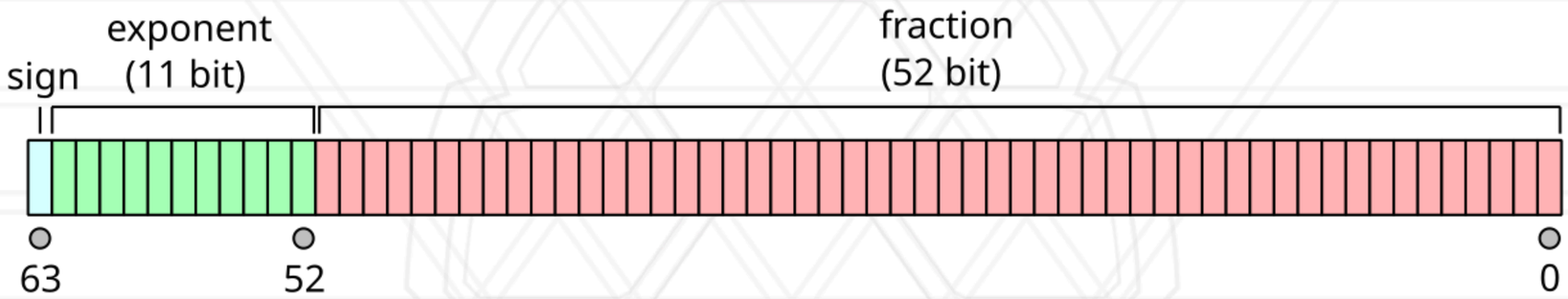
Quantization

Abhinav Bhatele, Department of Computer Science

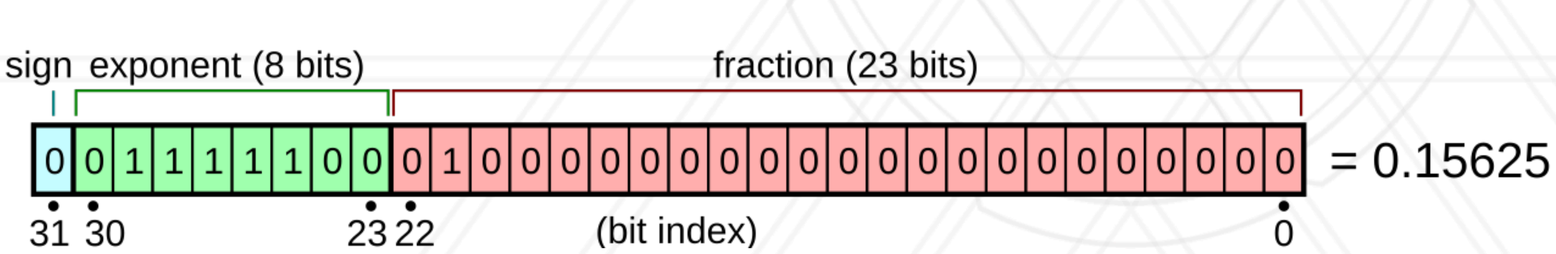
UNIVERSITY OF
MARYLAND

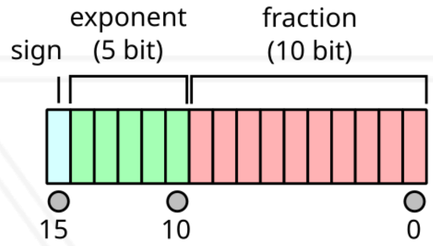
Floating point numbers

- Binary64 or double-precision (FP64) gives 15 to 17 significant digits of decimal precision



- In deep learning, we may not need all this precision
  - Binary32 (single-precision, FP32) or binary16 (half-precision, FP16, BF16) might be sufficient





https://en.wikipedia.org/wiki/IEEE_754

Quantization

- Map weights/activations/other states from higher to lower precision
  - FP32 to FP16, FP16 to FP8, …
- Why quantize?
  - Save memory
  - Potentially faster computation
- When not to quantize
  - Accuracy suffers
  - Performance suffers (if not supported in hardware natively)

Native hardware support

- FP16: NVIDIA Volta (V100)
- FP8: NVIDIA Hopper (H100, H200)
- FP4 GEMM: NVIDIA Blackwell (B100, GB200)

https://docs.nvidia.com/deeplearning/tensorrt-rtx/latest/getting-started/support-matrix-1/1.4.html

Types of quantization

- Post-training quantization
  - Quantize the weights of a pre-trained model
  - Simple and effective
- Quantization-aware training
  - More training complexity and potentially computation costs

Layer-wise quantization

- Goal: find a matrix of quantized weights that minimizes the squared error compared to full precision

$$\text{argmin}_{\widehat{\mathbf{W}}} ||\mathbf{W}\mathbf{X} - \widehat{\mathbf{W}}\mathbf{X}||_2^2$$

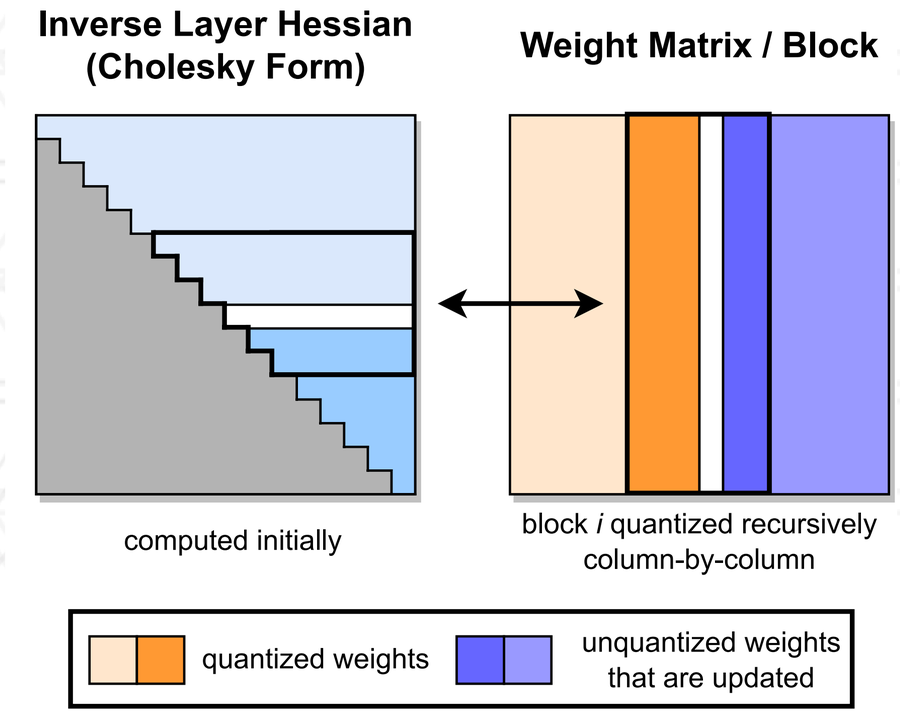
- Optimal Brain Quantization (OBQ): write equation above as sum of the quantized errors over each row of W

$$w_q = \text{argmin}_{w_q} \frac{(\text{quant}(w_q) - w_q)^2}{[\mathbf{H}_F^{-1}]_{qq}}, \quad \boldsymbol{\delta}_F = -\frac{w_q - \text{quant}(w_q)}{[\mathbf{H}_F^{-1}]_{qq}} \cdot (\mathbf{H}_F^{-1})_{:,q}.$$

- Quantize one weight at a time while updating all not-yet-quantized weights

GPTQ: Post-training quantization

- Arbitrary order: OBQ quantizes weights in greedy order
- **Fixed order:** Quantize weights in fixed order (all rows in the same order)
- Lazy batch updates
- Cholesky reformulation



UNIVERSITY OF
MARYLAND

Abhinav Bhatele

5218 Brendan Iribe Center (IRB) / College Park, MD 20742

phone: 301.405.4507 / e-mail: bhatele@cs.umd.edu