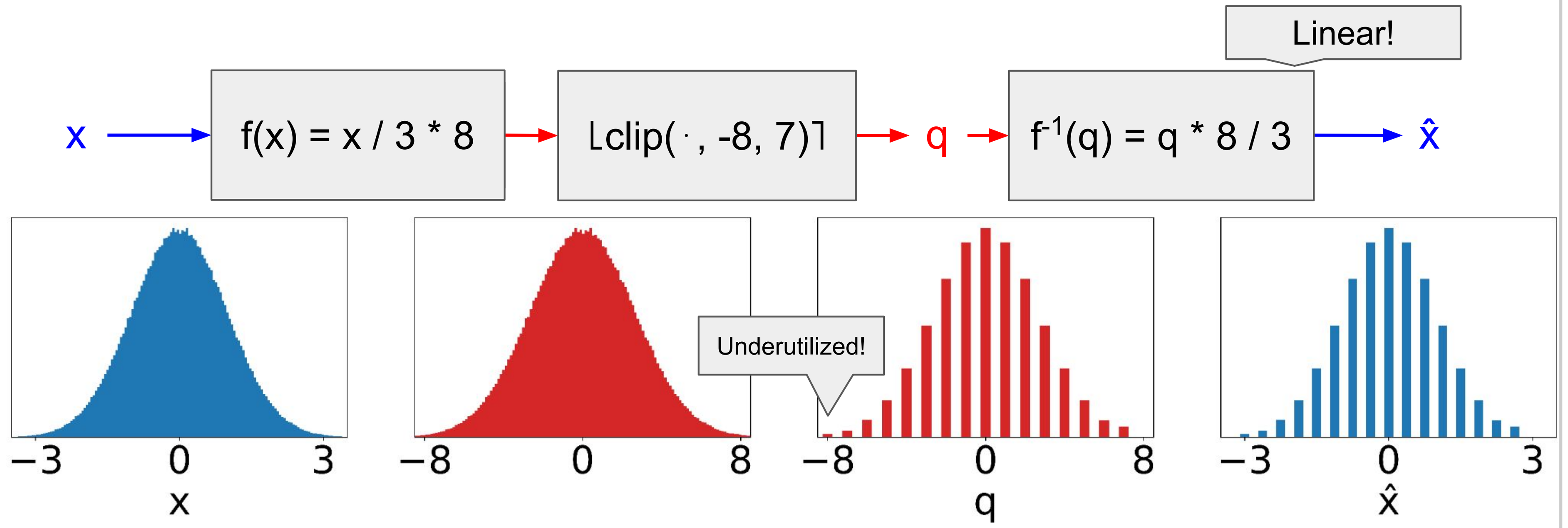
Boosting Entropy with Bell Box Quantization

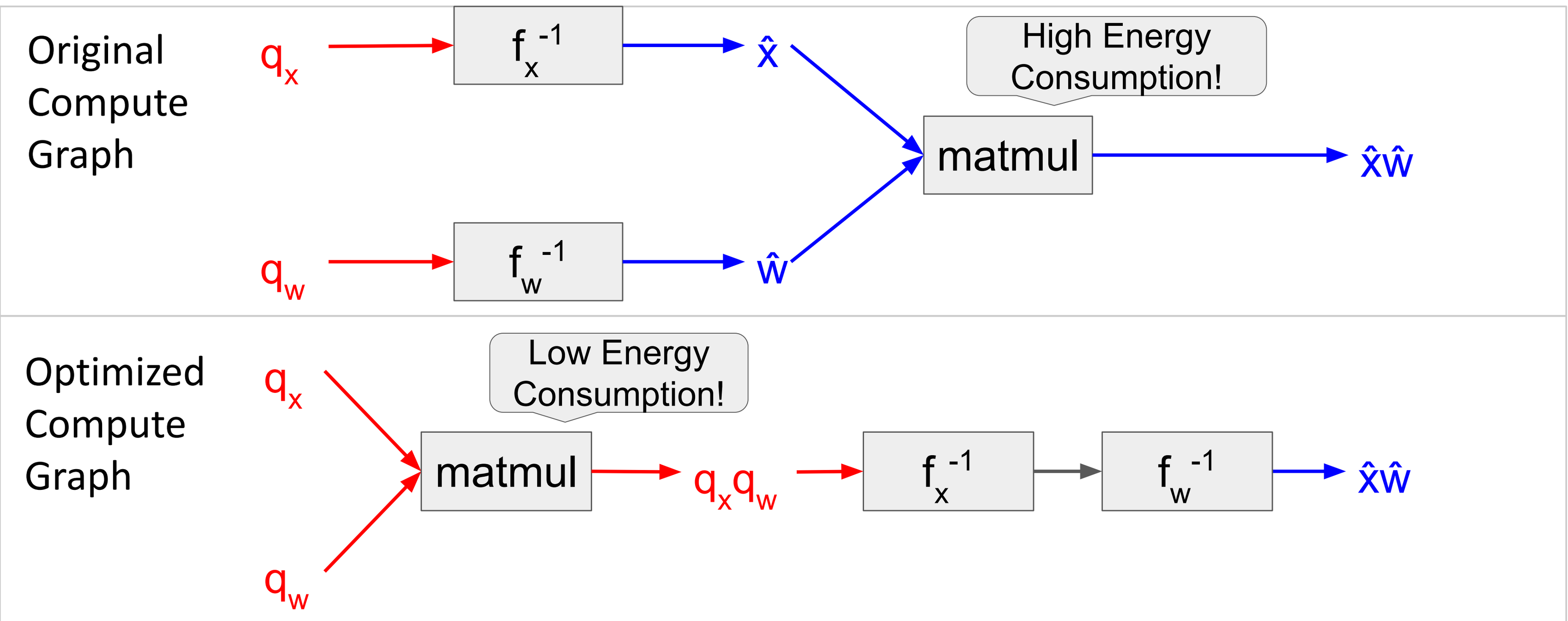
Ningfeng Yang and Tor M. Aamodt

Summary

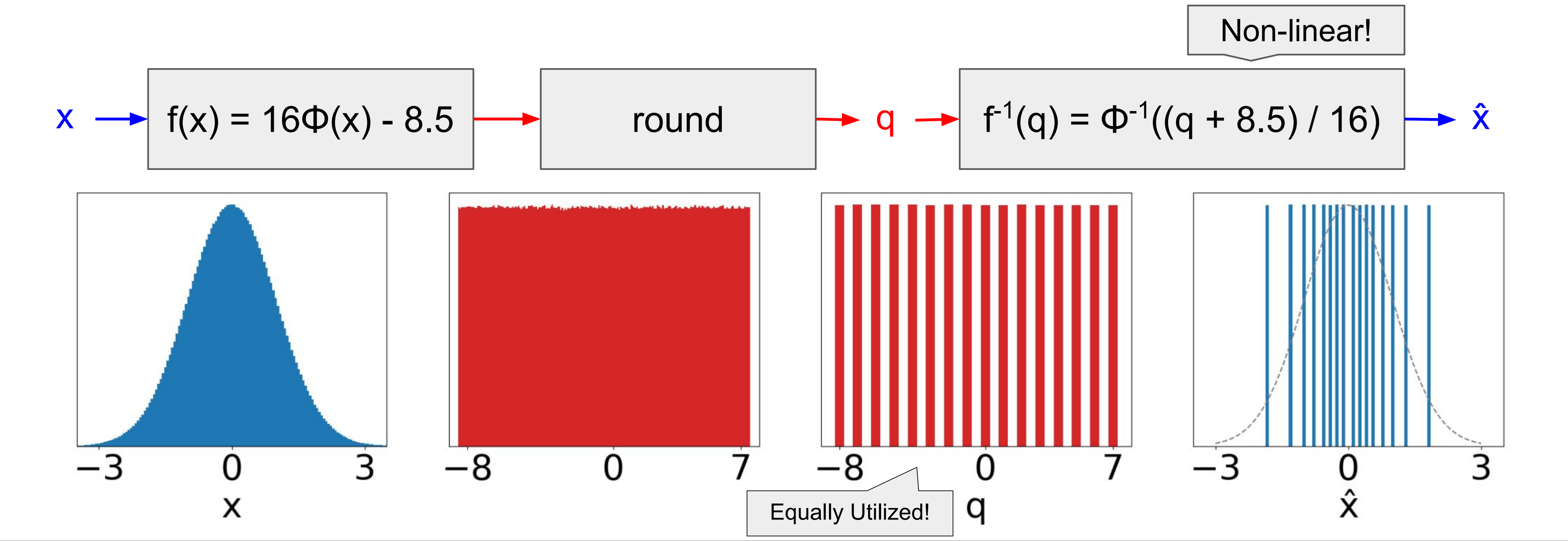
- Deep Neural Network inference is energy-intensive
  - Challenging for energy-constrained edge devices
    - Smartwatches, smartglasses, smartphones, robots, and drones
- Quantization-aware Pre-Training (QAPT) can
  - Reduce complexity of DNNs while mitigating accuracy degradation
- Existing compute-efficient QAPT methods can reduce
  - Model size, inference latency, and energy consumption
  - But underutilize the model's ability to store information
- Existing information-theoretically optimal (ITO) quantization methods
  - e.g. Quantile Quantization
  - Can reduce this underutilization but are not compute efficient
- This work: Bell Box Quantization
  - The first compute-efficient ITO quantization method

Background: Compute-Efficient Quantization Methods

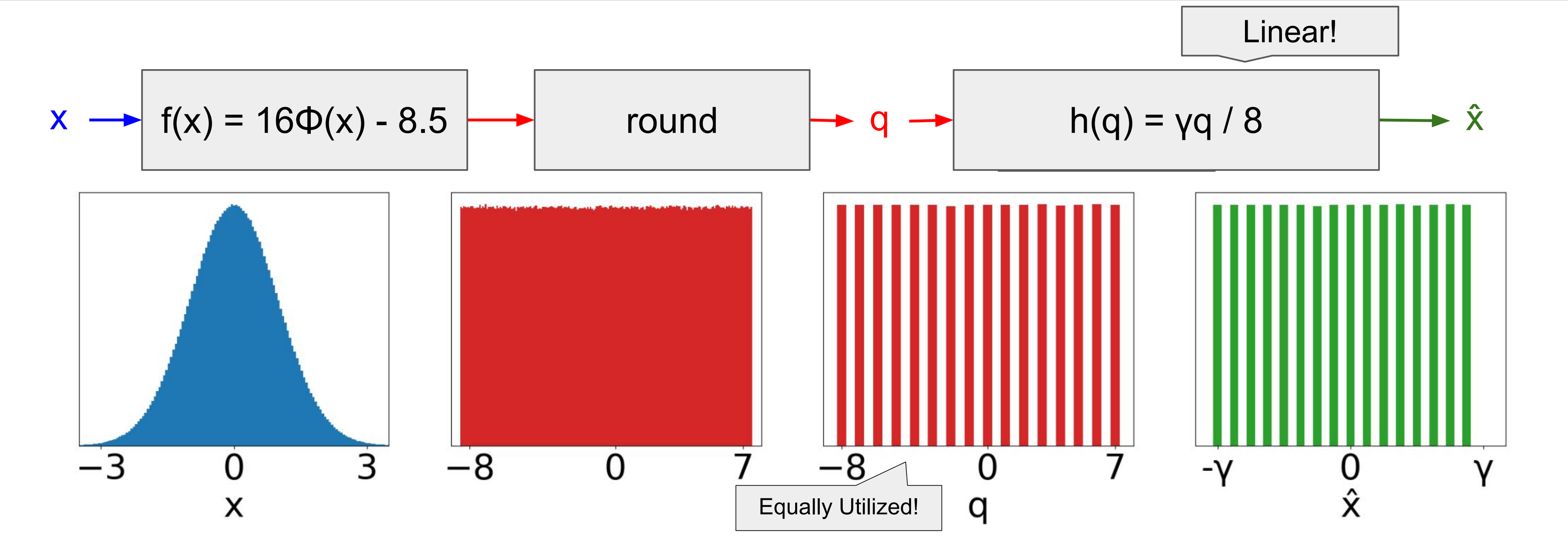




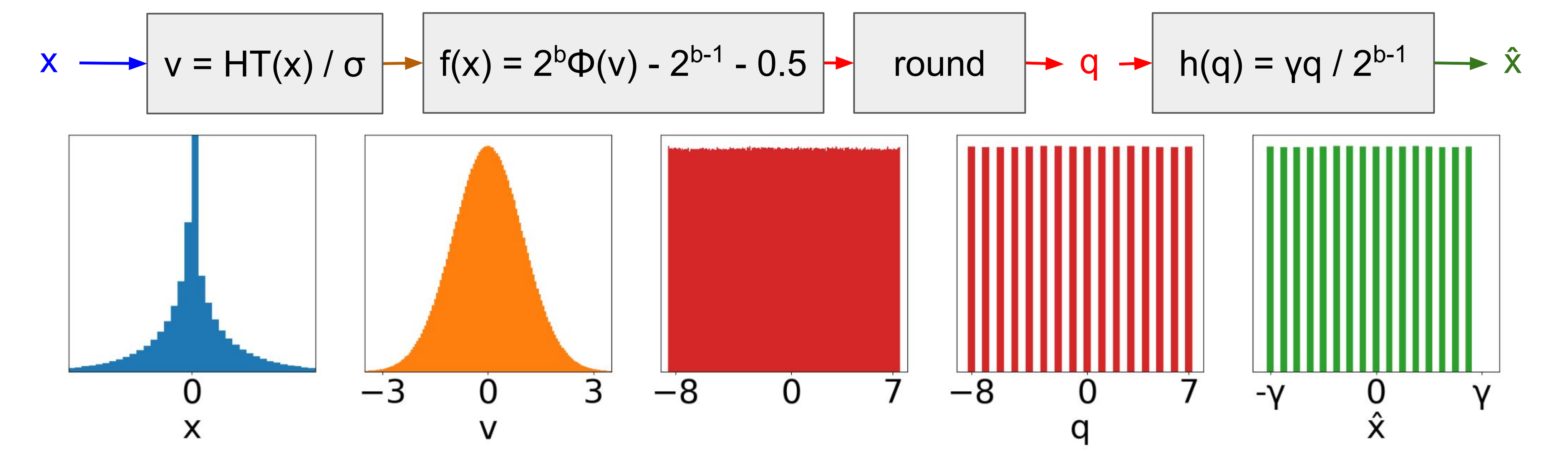
Background: Quantile Quantization



This Work: Equal Utilization and Linear Dequantization



Full Method: Bell Box Quantization



Experiments

Params	Tokens	Full	Bits	BBQ		QuEST		LSQ	
				Entropy	Perplexity	Entropy	Perplexity	Entropy	Perplexity
95M	3B	**24.75**	4	**3.93**	**25.51**	3.61	26.37	3.59	27.46
95M	3B	**24.75**	3	**2.96**	**26.55**	2.78	29.04	2.74	30.27
95M	3B	**24.75**	2	**1.97**	**31.34**	1.92	35.58	1.69	36.58
95M	3B	**24.75**	1	**1**	**49.22**	**1**	67.78	-	-
125M	5B	**21.51**	4	**3.93**	**22.15**	3.61	22.98	3.6	23.77
125M	5B	**21.51**	3	**2.96**	**23.22**	2.78	25.21	2.74	26.28
125M	5B	**21.51**	2	**1.98**	**27.34**	1.93	31.32	1.81	31.42
125M	5B	**21.51**	1	**1**	**44.58**	**1**	72.54	-	-
200M	10B	**17.93**	4	**3.93**	**18.79**	3.61	19.06	2.73	1778
200M	10B	**17.93**	3	**2.96**	**19.74**	2.78	20.82	2.5	140.9
200M	10B	**17.93**	2	**1.98**	**23.08**	1.93	25.46	1.63	78.19
200M	10B	**17.93**	1	**1**	**38.27**	**1**	52.37	-	-
300M	20B	**15.43**	4	**3.93**	**16.1**	3.61	16.26	-	-
300M	20B	**15.43**	3	**2.96**	**16.9**	2.78	17.67	-	-
300M	20B	**15.43**	2	**1.98**	**19.75**	1.93	21.53	-	-

Table 1: LLaMA models pre-trained on the C4 dataset.

Params	95M	95M	95M	95M	95M	95M	95M	95M	95M	95M	95M	95M
Bits WA	16	4	4	4	4	3	3	3	3	3	1	1
Method	-	BBQ	QuEST	LSQ	NF4	BBQ	QuEST	LSQ	NF3	N2UQ	BBQ	QuEST
Perplex	25	**26.7**	26.83	27.55	28.5	**28.61**	29.7	30.85	36.16	29.87	**53.8**	77.96
Params	95M	95M	95M	95M	95M	95M	125M	125M	125M	125M	125M	125M
Bits WA	2	2	2	2	2	2	3	3	2	2	1	1
Method	BBQ	QuEST	LSQ	NF2	SEQ	N2UQ	BBQ	QuEST	BBQ	QuEST	BBQ	QuEST
Perplex	**34.36**	36.7	38.98	246.6	40.07	35.59	**24.78**	26.05	**29.38**	31.47	**50.46**	69.78

Table 2: GPT models pre-trained on the C4 dataset

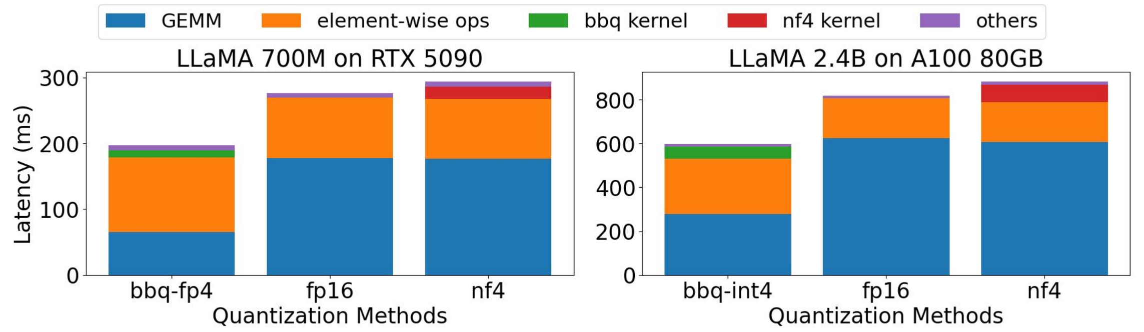


Figure 1: Profiling on NVIDIA GPUs

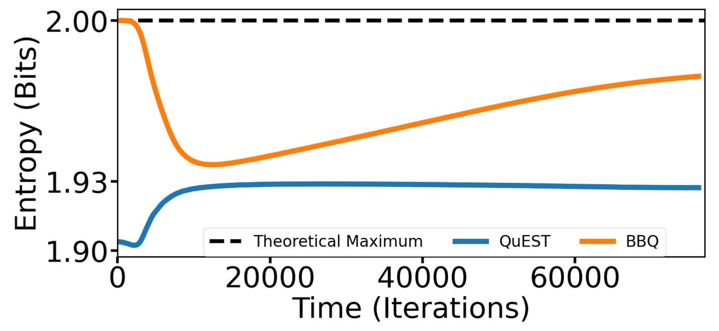


Figure 2: Entropy During Training

References:

Steven K. Esser, Jeffrey L. McKinstry, Deepika Bablani, Rathinakumar Appuswamy, & Dharmendra S. Modha (2020). LEARNED STEP SIZE QUANTIZATION. In International Conference on Learning Representations.

Andrei Panferov, Jiale Chen, Soroush Tabesh, Mahdi Nikdan, & Dan Alistarh (2025). QuEST: Stable Training of LLMs with 1-Bit Weights and Activations. In Forty-second International Conference on Machine Learning.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. Advances in Neural Information Processing Systems 36 (NeurIPS 2023)

Tim Dettmers, Mike Lewis, Sam Shleifer, Luke Zettlemoyer. 8-bit Optimizers via Block-wise Quantization. ICLR 2022.

Zechun Liu, Kwang-Ting Cheng, Dong Huang, Eric Xing, Zhiqiang Shen. Nonuniform-to-Uniform Quantization: Towards Accurate Quantization via Generalized Straight-Through Estimation. CVPR 2022